3 Probability and Statistical Distributions

"There are three kinds of lies: lies, damned lies, and statistics." (popularized by Mark Twain)
"In ancient times they had no statistics so they had to fall back on lies." (Stephen Leacock)

The main purpose of this chapter is to review notation and basic concepts in probability and statistics. The coverage of various topics cannot be complete, and it is aimed at concepts needed to understand material covered in the book. For an in-depth discussion of probability and statistics, please refer to numerous readily available textbooks, such as Bar89, Lup93, WJ03, Wass10, mentioned in §1.3.

The chapter starts with a brief overview of probability and random variables, then it reviews the most common univariate and multivariate distribution functions, and correlation coefficients. We also summarize the central limit theorem and discuss how to generate mock samples (random number generation) for a given distribution function.

Notation

Notation in probability and statistics is highly variable and ambiguous, and can make things confusing all on its own (and even more so in data mining publications!). We try to minimize the notational clutter, though this is not always possible. We have already introduced some notation in §1.2. For example, lowercase letters are used for probability density (differential distribution) functions (pdf), and the corresponding uppercase letter for their cumulative counterpart (cdf), for example, $h(x)$ and $H(x)$.

We have been able to simplify our nomenclature by ignoring some annoying difficulties, particularly in the case of continuous values. In reality, we cannot talk about the probability of x taking on a specific real-number value as being anything other than zero. The product $h(x)\,dx$ gives a probability that the value x would fall in a dx wide interval around x, but we will not explicitly write dx (see Wass10 for a clear treatment in this regard).

We shall use p for probability whenever possible, both for the probability of a single event and for the probability density functions (pdf).

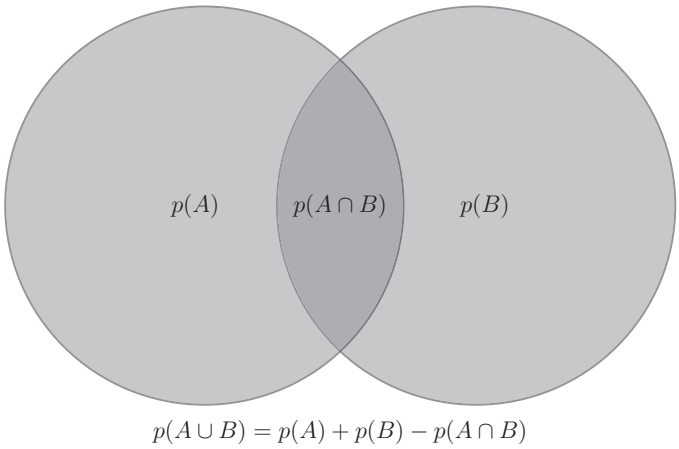

Figure 3.1. A representation of the sum of probabilities in eq. 3.1.

3.1. Brief Overview of Probability and Random Variables

3.1.1. Probability Axioms

Given an event A, such as the outcome of a coin toss, we assign it a real number $p(A)$, called the *probability* of A. As discussed above, $p(A)$ could also correspond to a probability that a value of x falls in a dx wide interval around x. To qualify as a probability, $p(A)$ must satisfy three Kolmogorov axioms:

1. $p(A) \geq 0$ for each A.
2. $p(\Omega) = 1$, where Ω is a set of all possible outcomes.
3. If A_1, A_2, ... are disjoint events, then $p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$, where $\bigcup$ stands for "union."

As a consequence of these axioms, several useful rules can be derived. The probability that the union of two events, A and B, will happen is given by the sum rule,

$$p(A \cup B) = p(A) + p(B) - p(A \cap B), \tag{3.1}$$

where $\cap$ stands for "intersection." That is, the probability that either A or B will happen is the sum of their respective probabilities minus the probability that *both* A and B will happen (this rule avoids the double counting of $p(A \cap B)$ and is easy to understand graphically: see figure 3.1).

If the complement of event A is $\overline{A}$, then

$$p(A) + p(\overline{A}) = 1. \tag{3.2}$$

The probability that both A and B will happen is equal to

$$p(A \cap B) = p(A|B)\, p(B) = p(B|A)\, p(A). \tag{3.3}$$

Here "|" is pronounced "given" and $p(A|B)$ is the probability of event A given that (conditional on) B is true. We discuss conditional probabilities in more detail in §3.1.3.

If events B_i, $i = 1, \ldots, N$ are disjoint and their union is the set of all possible outcomes, then

$$p(A) = \sum_i p(A \cap B_i) = \sum_i p(A|B_i)\, p(B_i). \tag{3.4}$$

This expression is known as the law of total probability. Conditional probabilities also satisfy the law of total probability. Assuming that an event C is not mutually exclusive with A or any of B_i, then

$$p(A|C) = \sum_i p(A|C \cap B_i)\, p(B_i|C). \tag{3.5}$$

Cox derived the same probability rules starting from a different set of axioms than Kolmogorov [2]. Cox's derivation is used to justify the so-called "logical" interpretation of probability and the use of Bayesian probability theory (for an illuminating discussion, see chapters 1 and 2 in Jay03). To eliminate possible confusion in later chapters, note that both the Kolmogorov and Cox axioms result in essentially the same probabilistic framework.

The difference between classical inference and Bayesian inference is fundamentally in the interpretation of the resulting probabilities (discussed in detail in chapters 4 and 5). Briefly, classical statistical inference is concerned with $p(A)$, interpreted as the long-term outcome, or frequency with which A occurs (or would occur) in identical repeats of an experiment, and events are restricted to propositions about random variables (see below). Bayesian inference is concerned with $p(A|B)$, interpreted as the plausibility of a proposition A, conditional on the truth of B, and A and B can be any logical proposition (i.e., they are not restricted to propositions about random variables).

3.1.2. Random Variables

A random, or stochastic, variable is, roughly speaking, a variable whose value results from the measurement of a quantity that is subject to random variations. Unlike normal mathematical variables, a random variable can take on a set of possible different values, each with an associated probability. It is customary in the statistics literature to use capital letters for random variables, and a lowercase letter for a particular realization of random variables (called random variates). We shall use lowercase letters for both.

There are two main types of random variables: discrete and continuous. The outcomes of discrete random variables form a countable set, while the outcomes of continuous random variables usually map on to the real number set (though one can define mapping to the complex plane, or use matrices instead of real numbers, etc.). The function which ascribes a probability value to each outcome of the random variable is the probability density function (pdf).

Independent identically distributed (iid) random variables are drawn from the same distribution and are independent. Two random variables, x and y, are *independent* if and only if

$$p(x, y) = p(x)\, p(y) \tag{3.6}$$

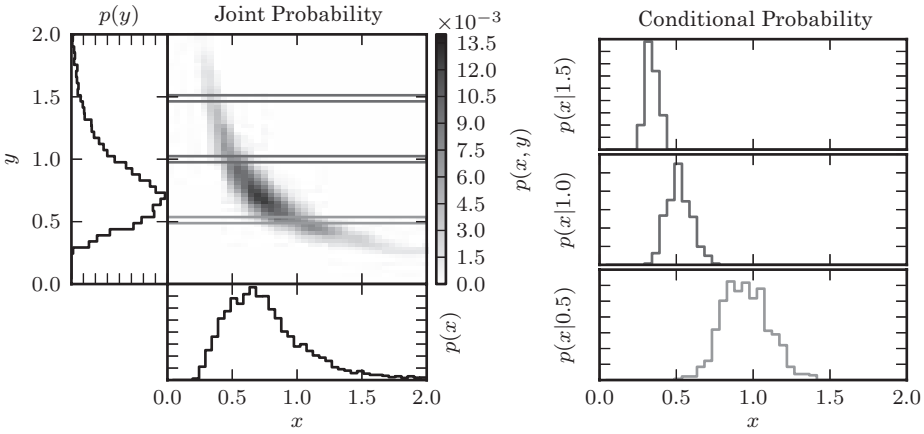

Figure 3.2. An example of a two-dimensional probability distribution. The color-coded panel shows $p(x, y)$. The two panels to the left and below show marginal distributions in x and y (see eq. 3.8). The three panels to the right show the conditional probability distributions $p(x|y)$ (see eq. 3.7) for three different values of y (as marked in the left panel).

for all values x and y. In other words, the knowledge of the value of x tells us nothing about the value of y.

The *data* are specific ("measured") values of random variables. We will refer to measured values as x_i, and to the set of all N measurements as $\{x_i\}$.

3.1.3. Conditional Probability and Bayes' Rule

When two continuous random variables are not independent, it follows from eq. 3.3 that

$$p(x, y) = p(x|y)\, p(y) = p(y|x)\, p(x). \tag{3.7}$$

The *marginal probability function* is defined as

$$p(x) = \int p(x, y)\, dy, \tag{3.8}$$

and analogously for $p(y)$. Note that complete knowledge of the conditional pdf $p(y|x)$, and the marginal probability $p(x)$, is sufficient to fully reconstruct $p(x, y)$ (the same is true with x and y reversed).

By combining eqs. 3.7 and 3.8, we get a continuous version of the law of total probability,

$$p(x) = \int p(x|y)p(y)\, dy. \tag{3.9}$$

An example of a two-dimensional probability distribution is shown in figure 3.2, together with corresponding marginal and conditional probability distributions. Note that the conditional probability distributions $p(x|y = y_0)$ are simply one-dimensional "slices" through the two-dimensional image $p(x, y)$ at given values

of y_0, and then divided (renormalized) by the value of the marginal distribution $p(y)$ at $y = y_0$. As a result of this renormalization, the integral of $p(x|y)$ (over x) is unity.

Eqs. 3.7 and 3.9 can be combined to yield *Bayes' rule*:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)\,dy}. \tag{3.10}$$

Bayes' rule relates conditional and marginal probabilities to each other. In the case of a discrete random variable, y_j, with M possible values, the integral in eq. 3.10 becomes a sum:

$$p(y_j|x) = \frac{p(x|y_j)p(y_j)}{p(x)} = \frac{p(x|y_j)p(y_j)}{\sum_{j=1}^{M} p(x|y_j)p(y_j)}. \tag{3.11}$$

Bayes' rule follows from a straightforward application of the rules of probability and is by no means controversial. It represents the foundation of Bayesian statistics, which has been a very controversial subject until recently. We briefly note here that it is not the rule itself that has caused controversy, but rather its application. Bayesian methods are discussed in detail in chapter 5.

We shall illustrate the use of marginal and conditional probabilities, and of Bayes' rule, with a simple example.

Example: the Monty Hall problem

The following problem illustrates how different probabilistic inferences can be derived about the same physical system depending on the available prior information. There are N=1000 boxes, of which 999 are empty and one contains some "prize." You choose a box at random; the probability that it contains the prize is 1/1000. This box remains closed. The probability that any one of other 999 boxes contains the prize is also 1/1000, and the probability that the box with the prize is among those 999 boxes is 999/1000. Then another person who *knows which box contains the prize* opens 998 empty boxes chosen from the 999 remaining boxes (i.e., the box you chose is "set aside"). It is important to emphasize that these 998 boxes are *not* selected randomly from the set of 999 boxes you did not choose—instead, they are selected as empty boxes. So, the remaining 999th box is almost certain to contain the prize; the probability is 999/1000 because there is a chance of only 1 in 1000 that the prize is in the box you chose initially, and the probabilities for the two unopened boxes must add up to 1. Alternatively, before 998 empty boxes were opened, the probability that the 999 boxes contained the prize was 999/1000. Given that all but one were demonstrated to be empty, the last 999th box now contains the prize with the same probability. If you were offered to switch the box you initially chose with other unopened box, you would increase the chances of getting the prize by a factor of 999 (from 1/1000 to 999/1000). On the other hand, if a third person walked in and had to choose one of the two remaining unopened boxes, but *without knowing* that initially there were 1000 boxes, nor which one you initially chose, he or she would pick the box with the prize with a probability of 1/2. The difference in expected outcomes is due to different prior information, and it nicely illustrates that the probabilities we assign to events reflect the state of our knowledge.

This problem, first discussed in a slightly different form in 1959 by Martin Gardner in his "Mathematical Games" column [3] in *Scientific American* (the "Three Prisoner Problem"), sounds nearly trivial and uncontroversial. Nevertheless, when the same mathematical problem was publicized for the case of $N = 3$ by Marilyn vos Savant in her newspaper column in 1990 [6], it generated an amazing amount of controversy. Here is a transcript of her column:

> Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?

vos Savant also provided the correct answer to her question (also known as "the Monty Hall problem"): you should switch the doors because it increases your chance of getting the car from 1/3 to 2/3. After her column was published, vos Savant received thousands of letters from readers, including many academics and mathematicians,[1] all claiming that vos Savant's answer is wrong and that the probability is 1/2 for both unopened doors. But as we know from the less confusing case with large N discussed above (this is why we started with the $N = 1000$ version), vos Savant was right and the unhappy readers were wrong. Nevertheless, if you side with her readers, you may wish to write a little computer simulation of this game and you will change your mind (as did about half of her readers). Indeed, vos Savant called on math teachers to perform experiments with playing cards in their classrooms—they *experimentally* verified that it pays to switch! Subsequently, the problem was featured in a 2011 episode of the pop-science television series *MythBusters*, where the hosts reached the same conclusion.

Here is a formal derivation of the solution using Bayes' rule. H_i is the hypothesis that the prize is in the ith box, and $p(H_i|I) = 1/N$ is its prior probability given background information I. Without a loss of generality, the box chosen initially can be enumerated as the first box. The "data" that $N - 2$ boxes, all but the first box and the kth box ($k > 1$), are empty is d_k (i.e., d_k says that the kth box remains closed). The probability that the prize is in the first box, given I and k, can be evaluated using Bayes' rule (see eq. 3.11),

$$p(H_1|d_k, I) = \frac{p(d_k|H_1, I)p(H_1|I)}{p(d_k|I)}. \tag{3.12}$$

The probability that the kth box remains unopened given that H_1 is true, $p(d_k|H_1, I)$, is $1/(N-1)$ because this box is randomly chosen from $N-1$ boxes. The denominator can be expanded using the law of total probability,

$$p(d_k|I) = \sum_{i=1}^{N} p(d_k|H_i, I)p(H_i|I). \tag{3.13}$$

[1] For amusing reading, check out http://www.marilynvossavant.com/articles/gameshow.html

The probability that the kth box stays unopened, given that H_i is true, is

$$p(d_k|H_i, I) = \begin{cases} 1 \text{ for } k = i, \\ 0 \text{ otherwise,} \end{cases} \tag{3.14}$$

except when $i = 1$ (see above). This reduces the sum to only two terms:

$$p(d_k|I) = p(d_k|H_1, I)p(H_1|I) + p(d_k|H_k, I)p(H_k|I) = \frac{1}{(N-1)N} + \frac{1}{N} = \frac{1}{N-1}. \tag{3.15}$$

This result might appear to agree with our intuition because there are $N - 1$ ways to choose one box out of $N - 1$ boxes, but this interpretation is not correct: the kth box is not chosen randomly in the case when the prize is not in the first box, but instead it *must* be chosen (the second term in the sum above). Hence, from eq. 3.12, the probability that the prize is in the first (initially chosen) box is

$$p(H_1|d_k, I) = \frac{\frac{1}{(N-1)}\frac{1}{N}}{\frac{1}{(N-1)}} = \frac{1}{N}. \tag{3.16}$$

It is easy to show that $p(H_k|d_k, I) = (N - 1)/N$. Note that $p(H_1|d_k, I)$ is equal to the prior probability $p(H_1|I)$; that is, the opening of $N - 2$ empty boxes (data d_k) did not improve our knowledge of the content of the first box (but it did improve our knowledge of the content of the kth box by a factor of $N - 1$).

Example: 2×2 contingency table

A number of illustrative examples about the use of conditional probabilities exist for the simple case of two discrete variables that can have two different values, yielding four possible outcomes. We shall use a medical test here: one variable is the result of a test for some disease, T, and the test can be negative (0) or positive (1); the second variable is the health state of the patient, or the presence of disease D: the patient can have a disease (1) or not (0). There are four possible combinations in this sample space: $T = 0, D = 0$; $T = 0, D = 1$; $T = 1, D = 0$; and $T = 1, D = 1$. Let us assume that we know their probabilities. If the patient is healthy ($T = 0$), the probability for the test being positive (a false positive) is $p(T = 1|D = 0) = \epsilon_{fP}$, where ϵ_{fP} is (typically) a small number, and obviously $p(T = 0|D = 0) = 1-\epsilon_{fP}$. If the patient has the disease ($T = 1$), the probability for the test being negative (a false negative) is $p(T = 0|D = 1) = \epsilon_{fN}$, and $p(T = 1|D = 1) = 1 - \epsilon_{fN}$. For a visual summary, see figure 3.3. Let us assume that we also know that the prior probability (in the absence of any testing, for example, based on some large population studies unrelated to our test) for the disease in question is $p(D = 1) = \epsilon_D$, where ϵ_D is a small number (of course, $p(D = 0) = 1 - \epsilon_D$).

Assume now that our patient took the test and it came out positive ($T = 1$). What is the probability that our patient has contracted the disease, $p(D = 1|T = 1)$?

D \ T	0	1
0	$1 - \epsilon_{fP}$	ϵ_{fP}
1	ϵ_{fN}	$1 - \epsilon_{fN}$

Figure 3.3. A contingency table showing $p(T|D)$.

Using Bayes' rule (eq. 3.11), we have

$$p(D = 1|T = 1) = \frac{p(T = 1|D = 1)\, p(D = 1)}{p(T = 1|D = 0)\, p(D = 0) + p(T = 1|D = 1)\, p(D = 1)}, \tag{3.17}$$

and given our assumptions,

$$p(D = 1|T = 1) = \frac{\epsilon_D - \epsilon_{fN}\,\epsilon_D}{\epsilon_D + \epsilon_{fP} - [\epsilon_D\,(\epsilon_{fP} + \epsilon_{fN})]}. \tag{3.18}$$

For simplicity, let us ignore second-order terms since all ϵ parameters are presumably small, and thus

$$p(D = 1|T = 1) = \frac{\epsilon_D}{\epsilon_D + \epsilon_{fP}}. \tag{3.19}$$

This is an interesting result: we can only reliably diagnose a disease (i.e., $p(D = 1|T = 1) \sim 1$) if $\epsilon_{fP} \ll \epsilon_D$. For rare diseases, the test must have an exceedingly low false-positive rate! On the other hand, if $\epsilon_{fP} \gg \epsilon_D$, then $p(D = 1|T = 1) \sim \epsilon_D/\epsilon_{fP} \ll 1$ and the testing does not produce conclusive evidence. The false-negative rate is not quantitatively important as long as it is not much larger than the other two parameters. Therefore, when being tested, it is good to ask about the test's false-positive rate.

If this example is a bit confusing, consider a sample of 1000 tested people, with $\epsilon_D = 0.01$ and $\epsilon_{fP} = 0.02$. Of those 1000 people, we expect that 10 of them have the disease and all, assuming a small ϵ_{fN}, will have a positive test. However, an additional $\sim$20 people will be selected due to a false-positive result, and we will end up with a group of 30 people who tested positively. The chance to pick a person with the disease will thus be 1/3.

An identical computation applies to a jury deciding whether a DNA match is sufficient to declare a murder suspect guilty (with all the consequences of such a verdict). In order for a positive test outcome to represent conclusive evidence, the applied DNA test must have a false-positive rate much lower than the probability of randomly picking the true murderer on the street. The larger the effective community

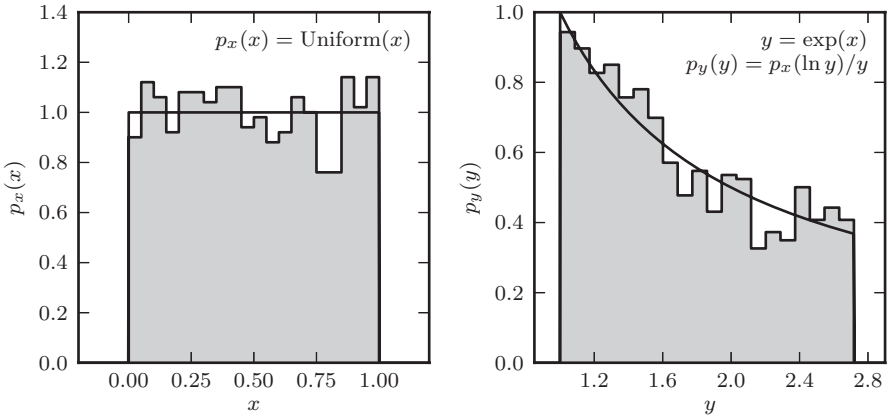

Figure 3.4. An example of transforming a uniform distribution. In the left panel, x is sampled from a uniform distribution of unit width centered on $x = 0.5$ ($\mu = 0$ and $W = 1$; see §3.3.1). In the right panel, the distribution is transformed via $y = \exp(x)$. The form of the resulting pdf is computed from eq. 3.20.

from which a murder suspect is taken (or DNA database), the better the DNA test must be to convincingly reach a guilty verdict.

These contingency tables are simple examples of the concepts which underlie *model selection* and *hypothesis testing*, which will be discussed in more detail in §4.6.

3.1.4. Transformations of Random Variables

Any function of a random variable is itself a random variable. It is a common case in practice that we measure the value of some variable x, but the interesting final result is a function $y(x)$. If we know the probability density distribution $p(x)$, where x is a random variable, what is the distribution $p(y)$, where $y = \Phi(x)$ (with $x = \Phi^{-1}(y)$)? It is easy to show that

$$p(y) = p\left[\Phi^{-1}(y)\right] \left| \frac{d\Phi^{-1}(y)}{dy} \right|. \tag{3.20}$$

For example, if $y = \Phi(x) = \exp(x)$, then $x = \Phi^{-1}(y) = \ln(y)$. If $p(x) = 1$ for $0 \le x \le 1$ and 0 otherwise (a uniform distribution), eq. 3.20 leads to $p(y) = 1/y$, with $1 \le y \le e$. That is, a uniform distribution of x is transformed into a nonuniform distribution of y (see figure 3.4).

Note that cumulative statistics, such as the median, do not change their order under monotonic transformations (e.g., given $\{x_i\}$, the median of x and the median of $\exp(x)$ correspond to the same data point).

If some value of x, say x_0, is determined with an uncertainty σ_x, then we can use a Taylor series expansion to estimate the uncertainty in y, say σ_y, at $y_0 = \Phi(x_0)$ as

$$\sigma_y = \left| \frac{d\Phi(x)}{dx} \right|_0 \sigma_x, \tag{3.21}$$

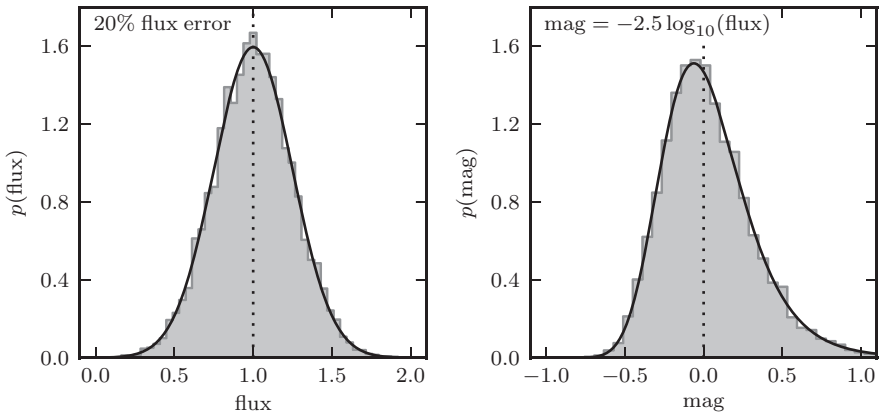

Figure 3.5. An example of Gaussian flux errors becoming non-Gaussian magnitude errors. The dotted line shows the location of the mean flux; note that this is not coincident with the peak of the magnitude distribution.

where the derivative is evaluated at x_0. While often used, this approach can produce misleading results when it is insufficient to keep only the first term in the Taylor series. For example, if the flux measurements follow a Gaussian distribution with a relative accuracy of a few percent, then the corresponding distribution of astronomical magnitudes (the logarithm of flux; see appendix C) is close to a Gaussian distribution. However, if the relative flux accuracy is 20% (corresponding to the so-called "5σ" detection limit), then the distribution of magnitudes is skewed and non-Gaussian (see figure 3.5). Furthermore, the mean magnitude is not equal to the logarithm of the mean flux (but the medians still correspond to each other!).

3.2. Descriptive Statistics

An arbitrary distribution function $h(x)$ can be characterized by its "location" parameters, "scale" or "width" parameters, and (typically dimensionless) "shape" parameters. As discussed below, these parameters, called descriptive statistics, can describe both various analytic distribution functions, as well as being determined directly from data (i.e., from our estimate of $h(x)$, which we named $f(x)$). When these parameters are based on $h(x)$, we talk about *population* statistics; when based on a finite-size data set, they are called *sample* statistics.

3.2.1. Definitions of Descriptive Statistics

Here are definitions for some of the more useful descriptive statistics:

- Arithmetic mean (also known as the expectation value),

$$\mu = E(x) = \int_{-\infty}^{\infty} x h(x)\, dx \tag{3.22}$$

- Variance,

$$V = \int_{-\infty}^{\infty} (x - \mu)^2 h(x)\, dx \tag{3.23}$$

- Standard deviation,

$$\sigma = \sqrt{V} \tag{3.24}$$

- Skewness,

$$\Sigma = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma} \right)^3 h(x)\, dx \tag{3.25}$$

- Kurtosis,

$$K = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma} \right)^4 h(x)\, dx \; - 3 \tag{3.26}$$

- Absolute deviation about d,

$$\delta = \int_{-\infty}^{\infty} |x - d| h(x)\, dx \tag{3.27}$$

- Mode (or the most probable value in case of unimodal functions), x_m,

$$\left(\frac{dh(x)}{dx} \right)_{x_m} = 0 \tag{3.28}$$

- p% quantiles (p is called a percentile), q_p,

$$\frac{p}{100} = \int_{-\infty}^{q_p} h(x)\, dx \tag{3.29}$$

Although this list may seem to contain (too) many quantities, remember that they are trying to capture the behavior of a completely general function $h(x)$. The variance, skewness, and kurtosis are related to the kth central moments (with $k = 2, 3, 4$) defined analogously to the variance (the variance is identical to the second central moment). The skewness and kurtosis are measures of the distribution shape, and will be discussed in more detail when introducing specific distributions below. Distributions that have a long tail toward x larger than the "central location" have positive skewness, and symmetric distributions have no skewness. The kurtosis is defined relative to the Gaussian distribution (thus it is adjusted by the "3" in eq. 3.26), with highly peaked ("leptokurtic") distributions having positive kurtosis, and flat-topped ("platykurtic") distributions

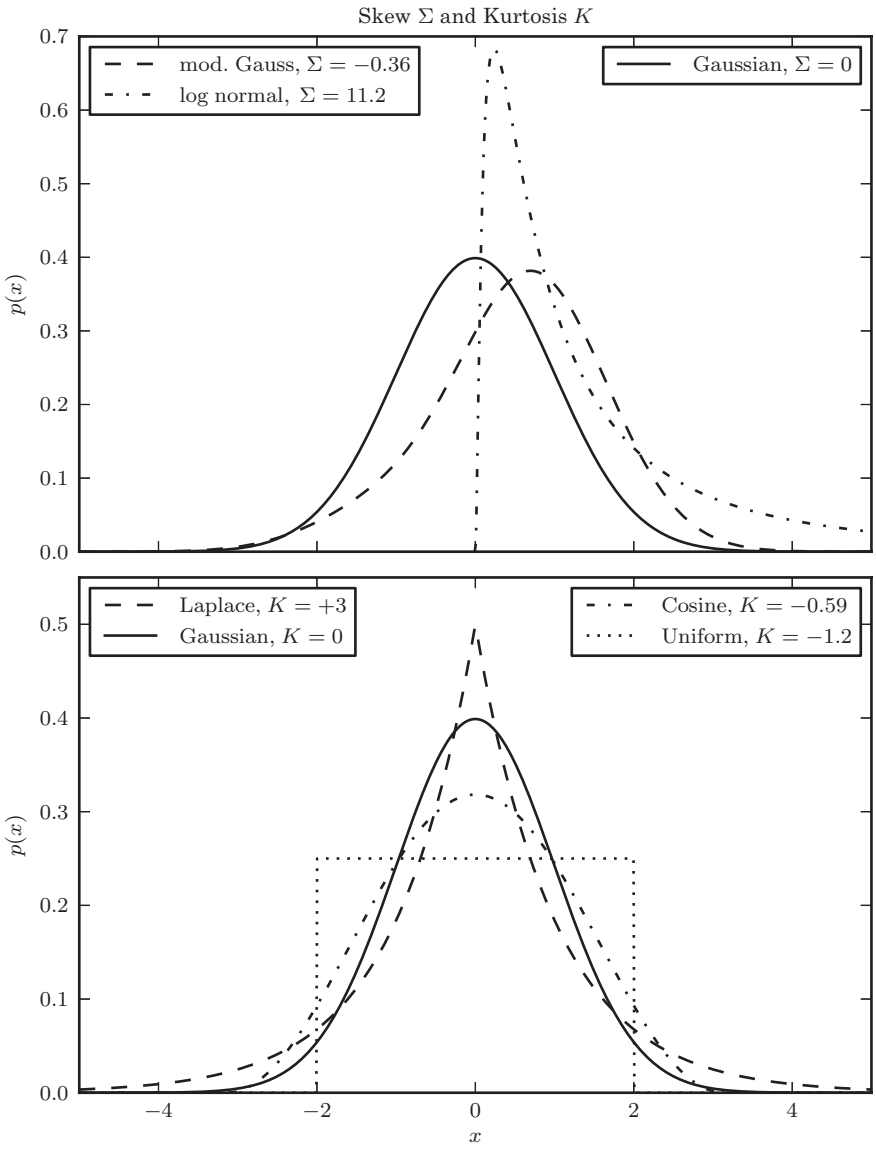

Figure 3.6. An example of distributions with different skewness Σ (top panel) and kurtosis K (bottom panel). The modified Gaussian in the upper panel is a normal distribution multiplied by a Gram–Charlier series (see eq. 4.70), with $a_0 = 2$, $a_1 = 1$, and $a_2 = 0.5$. The log-normal has $\sigma = 1.2$.

having negative kurtosis (see figure 3.6). The higher the distribution's moment, the harder it is to estimate it with small samples, and furthermore, there is more sensitivity to outliers (less robustness). For this reason, higher-order moments, such as skewness and kurtosis should be used with caution when samples are small.

The above statistical functions are among the many built into NumPy and SciPy. Useful functions to know about are `numpy.mean`, `numpy.median`, `numpy.var`, `numpy.percentile`, `numpy.std`, `scipy.stats.skew`, `scipy.stats.kurtosis`, and `scipy.stats.mode`. For example, to compute the quantiles of a one-dimensional array `x`, use the following:

```
import numpy as np
x = np.random.random(100) # 100 random numbers
q25, q50, q75 = np.percentile(x, [25, 50, 75])
```

For more information, see the NumPy and SciPy documentation of the above functions.

The absolute deviation about the mean (i.e., $d = \overline{x}$) is also called the mean deviation. When taken about the median, the absolute deviation is minimized. The most often used quantiles are the median, q_{50}, and the first and third quartile, q_{25} and q_{75}. The difference between the third and the first quartiles is called the interquartile range. A very useful relationship between the mode, the median and the mean, valid for mildly non-Gaussian distributions (see problem 2 in Lup93 for an elegant proof based on Gram–Charlier series[2]) is

$$x_m = 3\,q_{50} - 2\,\mu. \tag{3.30}$$

For example, this relationship is valid exactly for the Poisson distribution.

Note that some distributions do not have finite variance, such as the Cauchy distribution discussed below (§3.3.5). Obviously, when the distribution's variance is infinite (i.e., the tails of $h(x)$ do not decrease faster than x^{-3} for large $|x|$), the skewness and kurtosis will diverge as well.

3.2.2. Data-Based Estimates of Descriptive Statistics

Any of these quantities can be estimated directly from data, in which case they are called sample statistics (instead of population statistics). However, in this case we also need to be careful about the uncertainties of these estimates. Hereafter, assume that we are given N measurements, x_i, $i = 1, \ldots, N$, abbreviated as $\{x_i\}$. We will ignore for a moment the fact that measurements must have some uncertainty of their own (errors); alternatively, we can assume that x_i are measured much more accurately than the range of observed values (i.e., $f(x)$ reflects some "physics" rather than measurement errors). Of course, later in the book we shall relax this assumption.

In general, when estimating the above quantities for a sample of N measurements, the integral $\int_{-\infty}^{\infty} g(x)h(x)\,dx$ becomes proportional to the sum $\sum_i^N g(x_i)$, with the constant of proportionality $\sim(1/N)$. For example, the *sample arithmetic*

[2]The Gram–Charlier series is a convenient way to describe distribution functions that do not deviate strongly from a Gaussian distribution. The series is based on the product of a Gaussian distribution and the sum of the Hermite polynomials (see §4.7.4).

mean, $\overline{x}$, and the *sample standard deviation*, s, can be computed via standard formulas,

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.31}$$

and

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{3.32}$$

The reason for the $(N - 1)$ term instead of the naively expected N in the second expression is related to the fact that $\overline{x}$ is also determined from data (we discuss this subtle fact and the underlying statistical justification for the $(N - 1)$ term in more detail in §5.6.1). With N replaced by $N - 1$ (the so-called Bessel's correction), the sample variance (i.e., s^2) becomes unbiased (and the sample standard deviation given by expression 3.32 becomes a less biased, but on average still underestimated, estimator of the true standard deviation; for a Gaussian distribution, the underestimation varies from 20% for $N = 2$, to 3% for $N = 10$ and is less than 1% for $N > 30$). Similar factors that are just a bit different from N, and become N for large N, also appear when computing the skewness and kurtosis. What a "large N" means depends on a particular case and preset level of accuracy, but generally this transition occurs somewhere between $N = 10$ and $N = 100$ (in a different context, such as the definition of a "massive" data set, the transition may occur at N of the order of a million, or even a billion, again depending on the problem at hand).

We use different symbols in the above two equations ($\overline{x}$ and s) than in eqs. 3.22 and 3.24 (μ and σ) because the latter represent the "truth" (they are definitions based on the true $h(x)$, whatever it may be), and the former are simply *estimators* of that truth based on a *finite-size* sample ($\hat{x}$ is often used instead of $\overline{x}$). These estimators have a variance and a bias, and often they are judged by comparing their mean squared errors,

$$\mathrm{MSE} = V + \mathrm{bias}^2, \tag{3.33}$$

where V is the variance, and *the bias is defined as the expectation value of the difference between the estimator and its true (population) value*. Estimators whose variance and bias vanish as the sample size goes to infinity are called *consistent estimators*. An estimator can be unbiased but not consistent: as a simple example, consider taking the first measured value as an estimator of the mean value. This is unbiased, but its variance does not decrease with the sample size.

Obviously, we should also know the uncertainty in our estimators for μ ($\overline{x}$) and σ (s; note that s is *not* an uncertainty estimate for $\overline{x}$—this is a common misconception!). A detailed discussion of what exactly "uncertainty" means in this context, and how to derive the following expressions, can be found in chapter 5. Briefly, when N is large (at least 10 or so), and if the variance of $h(x)$ is finite, we expect from the central limit theorem (see below) that $\overline{x}$ and s will be distributed around their values given by eqs. 3.31 and 3.32 according to Gaussian distributions

with the widths (standard errors) equal to

$$\sigma_{\overline{x}} = \frac{s}{\sqrt{N}}, \tag{3.34}$$

which is called *the standard error of the mean*, and

$$\sigma_s = \frac{s}{\sqrt{2(N-1)}} = \frac{1}{\sqrt{2}}\sqrt{\frac{N}{N-1}}\,\sigma_{\overline{x}}. \tag{3.35}$$

The first expression is also valid when the standard deviation for parent population is known a priori (i.e., it is not determined from data using eq. 3.32). Note that for large N, the uncertainty of the location parameter is about 40% larger than the uncertainty of the scale parameter ($\sigma_{\overline{x}} \sim \sqrt{2}\,\sigma_s$). Note also that for small N, σ_s is not much smaller than s itself. The implication is that $s < 0$ is allowed according to the standard interpretation of "error bars" that implicitly assumes a Gaussian distribution! We shall return to this seemingly puzzling result in chapter 5 (§5.6.1), where an expression to be used instead of eq. 3.35 for small N (< 10) is derived.

Estimators can be compared in terms of their *efficiency*, which measures how large a sample is required to obtain a given accuracy. For example, the median determined from data drawn from a Gaussian distribution shows a scatter around the true location parameter (μ in eq. 1.4) larger by a factor of $\sqrt{\pi/2} \sim 1.253$ than the scatter of the mean value (see eq. 3.37 below). Since the scatter decreases with $1/\sqrt{N}$, the efficiency of the mean is $\pi/2$ times larger than the efficiency of the median. The smallest attainable variance for an unbiased estimator is called the *minimum variance bound (MVB)* and such an estimator is called the *minimum variance unbiased estimator (MVUE)*. We shall discuss in more detail how to determine the MVB in §4.2. Methods for estimating the bias and variance of various estimators are further discussed in §4.5 on bootstrap and jackknife methods. An estimator is *asymptotically normal* if its distribution around the true value approaches a Gaussian distribution for large sample size, with variance decreasing proportionally to $1/N$.

For the case of real data, which can have spurious measurement values (often, and hereafter, called "outliers"), quantiles offer a more robust method for determining location and scale parameters than the mean and standard deviation. For example, the median is a much more *robust* estimator of the location than the mean, and the interquartile range ($q_{75} - q_{25}$) is a more *robust* estimator of the scale parameter than the standard deviation. This means that the median and interquartile range are much less affected by the presence of outliers than the mean and standard deviation. It is easy to see why: if you take 25% of your measurements that are larger than q_{75} and arbitrarily modify them by adding a large number to all of them (or multiply them all by a large number, or different large numbers), both the mean and the standard deviation will be severely affected, while the median and the interquartile range will remain unchanged. Furthermore, even in the absence of outliers, for some distributions that do not have finite variance, such as the Cauchy distribution, the median and the interquartile range are the best choices for estimating location and scale parameters. Often, the interquartile range is renormalized so that the width estimator, σ_G, becomes an unbiased estimator of σ

for a *perfect*[3] Gaussian distribution (see §3.3.2 for the origin of the factor 0.7413),

$$\sigma_G = 0.7413\,(q_{75} - q_{25}). \tag{3.36}$$

There is, however, a price to pay for this robustness. For example, we already discussed that the efficiency of the median as a location estimator is poorer than that for the mean in the case of a Gaussian distribution. An additional downside is that it is much easier to compute the mean than the median for large samples; although the efficient algorithms described in §2.5.1 make this downside somewhat moot. In practice, one is often willing to pay the price of ∼25% larger errors for the median than for the mean (assuming nearly Gaussian distributions) to avoid the possibility of catastrophic failures due to outliers.

AstroML provides a convenience routine for calculating σ_G:

```
import numpy as np
from astroML import stats
x = np.random.normal(size=1000)  # 1000 normally
   # distributed points
stats.sigmaG(x)
1.0302378533978402
```

A very useful result is the following expression for computing standard error, σ_{qp}, for an arbitrary quantile q_p (valid for large N; see Lup93 for a derivation):

$$\sigma_{qp} = \frac{1}{h_p}\sqrt{\frac{p(1-p)}{N}}, \tag{3.37}$$

where h_p is the value of the probability distribution function at the pth percentile (e.g., for the median, $p = 0.5$). Unfortunately, σ_{qp} depends on the underlying $h(x)$. In the case of a Gaussian distribution, it is easy to derive that the standard error for the median is

$$\sigma_{q50} = s\sqrt{\frac{\pi}{2N}}, \tag{3.38}$$

with $h_{50} = 1/(s\sqrt{2\pi})$ and $s \sim \sigma$ in the limit of large N, as mentioned above. Similarly, the standard error for σ_G (eq. 3.36) is $1.06s/\sqrt{N}$, or about 50% larger than σ_s (eq. 3.35). The coefficient (1.06) is derived assuming that q_{25} and q_{75} are

[3] A real mathematician would probably laugh at placing the adjective *perfect* in front of "Gaussian" here. What we have in mind is a habit, especially common among astronomers, to (mis)use the word Gaussian for any distribution that even remotely resembles a bell curve, even when outliers are present. Our statement about the scatter of the median being larger than the scatter of the mean is not correct in such cases.

not correlated, which is not true for small samples (which can have scatter in σ_G up to 10–15% larger, even if drawn from a Gaussian distribution).

We proceed with an overview of the most common distributions used in data mining and machine learning applications. These distributions are called $p(x|I)$ to distinguish their mathematical definitions (equivalent to $h(x)$ introduced earlier) from $f(x)$ estimated from data. All the parameters that quantitatively describe the distribution $p(x)$ are "hidden" in I (for "information"), and $p(x|I)\,dx$ is interpreted as the probability of having a value of the random variable between x and $x + dx$. All $p(x|I)$ are normalized according to eq. 1.2.

3.3. Common Univariate Distribution Functions

Much of statistics is based on the analysis of distribution functions for random variables, and most textbooks cover this topic in great detail. From a large number of distributions introduced by statisticians, a relatively small number cover the lion's share of practical applications. We introduce them and provide their basic properties here, without pretending to cover all the relevant aspects. For an in-depth understanding, please consult statistics textbooks (e.g., Lup93, Wass10).

3.3.1. The Uniform Distribution

The uniform distribution is implemented in `scipy.stats.uniform`:

```
from scipy import stats
dist = stats.uniform(0, 2) # left edge = 0,
       # width = 2
r = dist.rvs(10)  # ten random draws
p = dist.pdf(1)  # pdf evaluated at x=1
```

The uniform distribution (also known as "top-hat" and "box") is described by

$$p(x|\mu, W) = \frac{1}{W} \text{ for } |x - \mu| \leq \frac{W}{2}, \tag{3.39}$$

and 0 otherwise, where W is the width of the "box" (see figure 3.7). Obviously, the mean and the median are equal to μ. Although simple, the uniform distribution often appears in practical problems (e.g., the distribution of the position of a photon detected in a finite-size pixel) and a well-known result is that

$$\sigma = \frac{W}{\sqrt{12}} \sim 0.3W, \tag{3.40}$$

where σ is computed using eq. 3.24. Since the uniform distribution is symmetric, its skewness is 0, and it is easy to show that its kurtosis is -1.2 (i.e., platykurtic as we would expect). Note that we can arbitrarily vary the location of this distribution

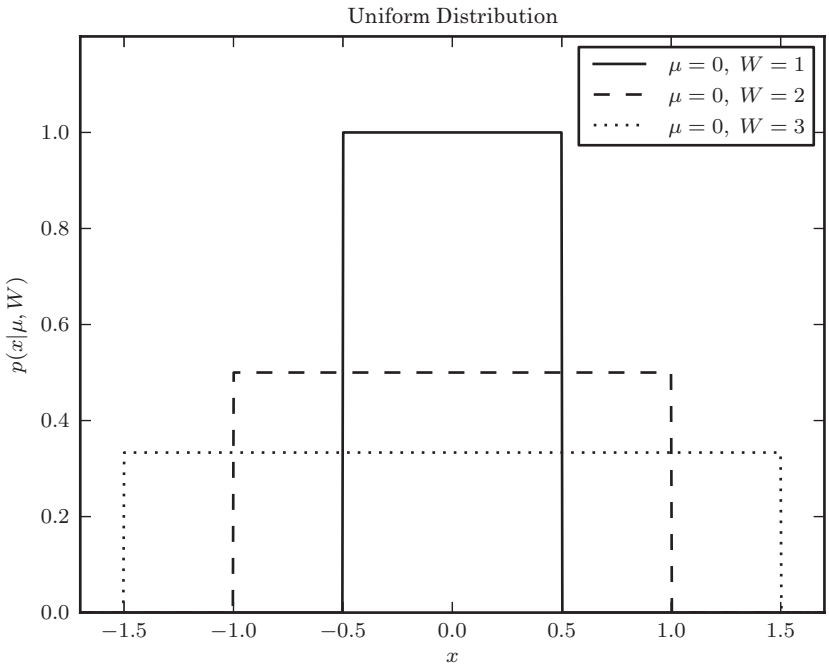

Figure 3.7. Examples of a uniform distribution (see eq. 3.39).

along the x-axis, and multiply x by an arbitrary scale factor, without any impact on the distribution's *shape*. That is, if we define

$$z = \frac{x - \mu}{W}, \tag{3.41}$$

then

$$p(z) = 1 \ \text{ for } |z| \leq \frac{1}{2}, \ \text{ and } \ p(z) = 0 \ \text{ for } |z| > \frac{1}{2}. \tag{3.42}$$

This independence of the distribution's shape on shift and normalization of the random variable x is a general result, and illustrates the meaning of the "location" and "scale" (or "width") parameters. An analogous "shift and rescale" transformation can be seen in the definitions of skewness and kurtosis (eqs. 3.25 and 3.26).

3.3.2. The Gaussian (normal) Distribution

The normal distribution is implemented in `scipy.stats.norm`:

```
from scipy import stats
dist = stats.norm(0, 1)  # mean = 0, stdev = 1
r = dist.rvs(10)  # ten random draws
p = dist.pdf(0) # pdf evaluated at x=0
```

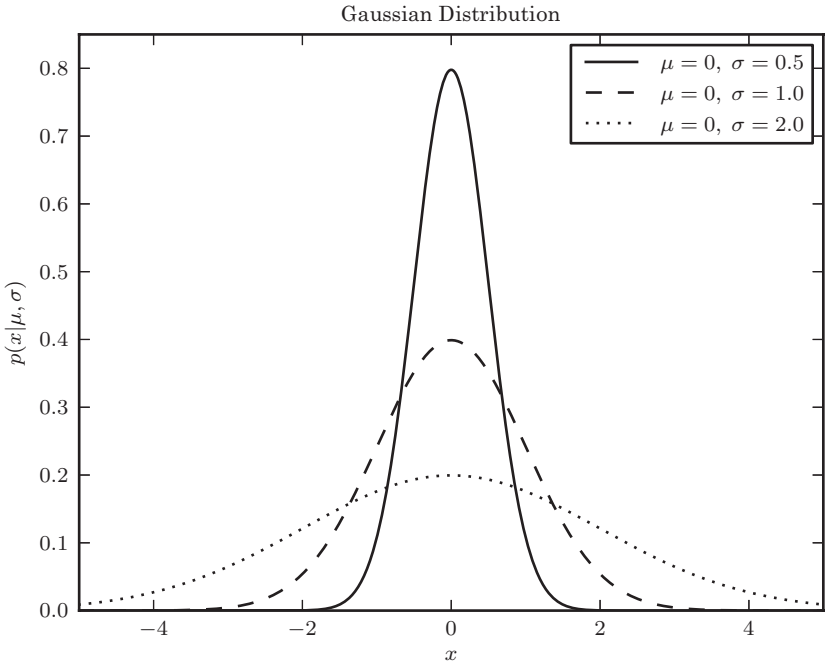

Figure 3.8. Examples of a Gaussian distribution (see eq. 3.43).

We have already introduced the Gaussian distribution in eq. 1.4, repeated here for completeness,

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right). \tag{3.43}$$

The Gaussian distribution is also called the normal distribution, $\mathcal{N}(\mu, \sigma)$. The standard deviation σ is often replaced by the variance σ^2 and the distribution is then referred to as $\mathcal{N}(\mu, \sigma^2)$, but here we uniformly adopt the first form. Since the Gaussian distribution is symmetric (see figure 3.8), its skewness is 0, and by the definition of kurtosis, its kurtosis is 0 as well (kurtosis is defined so that this is the case).

The Gaussian distribution has two main properties that make it special. First, it lends itself to analytic treatment in many cases; most notably, a convolution of two Gaussian distributions is also Gaussian (it's hard to believe, but computers haven't existed forever). The convolution of a function $f(x)$ with a function $g(x)$ (both assumed real functions) is defined as

$$(f \star g)(x) = \int_{-\infty}^{\infty} f(x')\, g(x - x')\, dx' = \int_{-\infty}^{\infty} f(x - x')\, g(x')\, dx'. \tag{3.44}$$

In particular, the convolution of a Gaussian distribution $\mathcal{N}(\mu_o, \sigma_o)$ (e.g., an intrinsic distribution we are trying to measure) with a Gaussian distribution $\mathcal{N}(b, \sigma_e)$ (i.e., Gaussian error distribution with bias b and random error σ_e) produces parameters

for the resulting Gaussian[4] $\mathcal{N}(\mu_C, \sigma_C)$ given by

$$\mu_C = (\mu_o + b) \quad \text{and} \quad \sigma_C = \left(\sigma_o^2 + \sigma_e^2\right)^{1/2}. \tag{3.45}$$

Similarly, the Fourier transform of a Gaussian is also a Gaussian (see §10.2.2). Another unique feature of the Gaussian distribution is that the sample mean and the sample variance are independent.

Second, the central limit theorem (see below) tells us that the mean of samples drawn from an almost arbitrary distribution will follow a Gaussian distribution. Hence, much of statistics and most classical results are based on an underlying assumption that a Gaussian distribution is applicable, although often it is not. For example, the ubiquitous least-squares method for fitting linear models to data is invalid when the measurement errors do not follow a Gaussian distribution (see chapters 5 and 8). Another interesting aspect of the Gaussian distribution is its "information content," which we discuss in chapter 5.

The cumulative distribution function for a Gaussian distribution,

$$P(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{-(x' - \mu)^2}{2\sigma^2}\right) dx', \tag{3.46}$$

cannot be evaluated in closed form in terms of elementary functions, and usually it is expressed in terms of the *Gauss error function*,

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left(-t^2\right) dt. \tag{3.47}$$

Tables and computer algorithms for evaluating erf(z) are readily available (note that erf(∞) = 1). With the aid of the error function,

$$P(x|\mu, \sigma) = \frac{1}{2}\left(1 \pm \text{erf}\left(\frac{|x - \mu|}{\sigma\sqrt{2}}\right)\right), \tag{3.48}$$

with the plus sign for $x > \mu$ and the minus sign otherwise.

The Gauss error function is available in `scipy.special`:

```
>>> from scipy.special import erf
>>> erf(1)
0.84270079294971478
```

[4]Note that the product of two Gaussians retains a Gaussian shape but its integral is *not* unity. Furthermore, the location and scale parameters are *very different* from those given by eq. 3.45. For example, the location parameter is a weighted sum of the two input location parameters (see §5.2.1 and eq. 5.20 for a detailed discussion).

Note that the integral of $p(x|\mu, \sigma)$ given by eq. 3.43 between two arbitrary integration limits, a and b, can be obtained as the difference of the two integrals $P(b|\mu, \sigma)$ and $P(a|\mu, \sigma)$. As a special case, the integral for $a = \mu - M\sigma$ and $b = \mu + M\sigma$ ("$\pm M\sigma$" ranges around μ) is equal to $\mathrm{erf}(M/\sqrt{2})$. The values for $M = 1$, 2, and 3 are 0.682, 0.954, and 0.997. The incidence level of "one in a million" corresponds to $M = 4.9$ and "one in a billion" to $M = 6.1$. Similarly, the interquartile range for the Gaussian distribution can be expressed as

$$q_{75} - q_{25} = \sigma\, 2\sqrt{2}\, \mathrm{erf}^{-1}(0.5) \approx 1.349\,\sigma, \tag{3.49}$$

which explains the coefficient in eq. 3.36.

If x follows a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, then $y = \exp(x)$ has a log-normal distribution. The log-normal distribution arises when the variable is a product of many independent positive variables (and the central limit theorem is applicable to the sum of their logarithms). The mean value of y is $\exp(\mu + \sigma^2/2)$, the mode is $\exp(\mu - \sigma^2)$, and the median is $\exp(\mu)$. Note that the mean value of y is *not* equal to $\exp(\mu)$, with the discrepancy increasing with the σ^2/μ ratio; this fact is related to our discussion of the failing Taylor series expansion at the end of §3.1.4.

3.3.3. The Binomial Distribution

The binomial distribution is implemented in `scipy.stats.binomial`:

```
from scipy import stats
dist = stats.binom(20, 0.7)  # N = 20, b = 0.7
r = dist.rvs(10)  # ten random draws
p = dist.pmf(8) # prob. evaluated at k=8
```

Unlike the Gaussian distribution, which describes the distribution of a continuous variable, the binomial distribution describes the distribution of a variable that can take only two discrete values (say, 0 or 1, or success vs. failure, or an event happening or not). If the probability of success is b, then the distribution of a discrete variable k (an integer number, unlike x which is a real number) that measures how many times success occurred in N trials (i.e., measurements), is given by

$$p(k|b, N) = \frac{N!}{k!(N-k)!}\, b^k\, (1-b)^{N-k} \tag{3.50}$$

(see figure 3.9). The special case of $N = 1$ is also known as a Bernoulli distribution. This result can be understood as the probability of k consecutive successes followed by $(N - k)$ consecutive failures (the last two terms), multiplied by the number of different permutations of such a draw (the first term). The mean (i.e., the expected number of successes) for the binomial distribution is $\overline{k} = bN$, and its standard

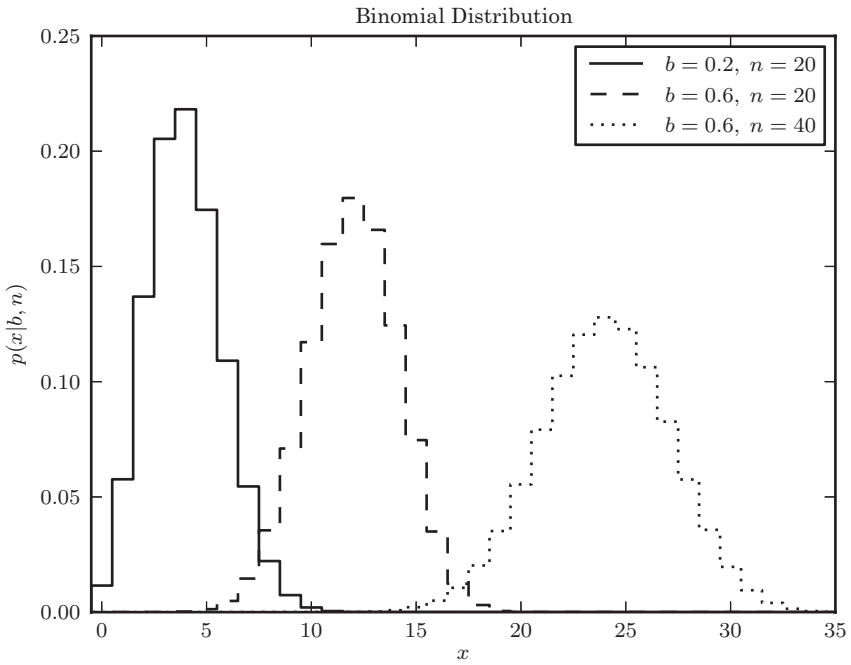

Figure 3.9. Examples of a binomial distribution (see eq. 3.50).

deviation is

$$\sigma_k = [N\, b\, (1 - b)]^{1/2}\,. \tag{3.51}$$

A common example of a process following a binomial distribution is the flipping of a coin. If the coin is fair, then $b = 0.5$, and success can be defined as obtaining a head (or a tail). For a real coin tossed N times, with k successes, we can ask what is our best estimate of b, called b^0, and its uncertainty given these data, σ_b. If N is large, we can simply determine $b^0 = k/N$, with its uncertainty (standard error) following from eq. 3.51 ($\sigma_b = \sigma_k/N$), and *assume* that the probability distribution for the true value of b is given by a Gaussian distribution $\mathcal{N}(b^0, \sigma_b)$. For example, it is easy to compute that it takes as many as 10^4 tosses of a fair coin to convince yourself that it is indeed fair within 1% accuracy ($b = 0.500 \pm 0.005$). How to solve this problem in a general case, without having to assume a Gaussian error distribution, will be discussed in chapter 5.

Coin flips are assumed to be independent events (i.e., the outcome of one toss does not influence the outcome of another toss). In the case of drawing a ball from a bucket full of red and white balls, we will get a binomial distribution if the bucket is infinitely large (so-called drawing with replacement). If instead the bucket is of finite size, then $p(k|b, N)$ is described by a hypergeometric distribution, which we will not use here.

A related distribution is the multinomial distribution, which is a generalization of the binomial distribution. It describes the distribution of a variable that can have more than two discrete values, say M values, and the probability of each value is

different and given by b_m, $b = 1, \ldots, M$. The multinomial distribution describes the distribution of M discrete variables k_m which count how many times the value indexed by m occurred in n trials.

The function `multinomial` in the `numpy.random` submodule implements random draws from a multinomial distribution:

```
from numpy.random import multinomial
vals = multinomial(10, pvals=[0.2, 0.3, 0.5])
# pvals sum to 1
```

3.3.4. The Poisson Distribution

The Poisson distribution is implemented in `scipy.stats.poisson`:

```
from scipy import stats
dist = stats.poisson(5)  # mu = 5
r = dist.rvs(10)  # ten random draws
p = dist.pmf(3) # prob. evaluated at k=3
```

The Poisson distribution is a special case of the binomial distribution and thus it also describes the distribution of a discrete variable. The classic example of this distribution, and an early application, is analysis of the chance of a Prussian cavalryman being killed by the kick of a horse. If the number of trials, N, for a binomial distribution goes to infinity such that the probability of success, $p = k/N$, stays fixed, then the distribution of the number of successes, k, is controlled by $\mu = pN$ and given by

$$p(k|\mu) = \frac{\mu^k \exp(-\mu)}{k!} \tag{3.52}$$

(see figure 3.10). The mean (or expectation) value is μ, and it fully describes a Poisson distribution: the mode (most probable value) is $(\mu - 1)$, the standard deviation is $\sqrt{\mu}$, the skewness is $1/\sqrt{\mu}$, and the kurtosis is $1/\mu$. As μ increases, both the skewness and kurtosis decrease, and thus the Poisson distribution becomes more and more similar to a Gaussian distribution, $\mathcal{N}(\mu, \sqrt{\mu})$ (and remember that the Poisson distribution is a limiting case of the binomial distribution). Interestingly, although the Poisson distribution morphs into a Gaussian distribution for large μ, the expectation value of the difference between the mean and median does *not* become 0, but rather 1/6. Because of this transition to a Gaussian for large μ, the Poisson distribution is sometimes called the "law of small numbers." Sometimes it is also called the "law of rare events" but we emphasize that μ need *not* be a small number—the adjective "rare" comes from the fact that only a small fraction of a large number of trials N results in success ("p is small, not μ").

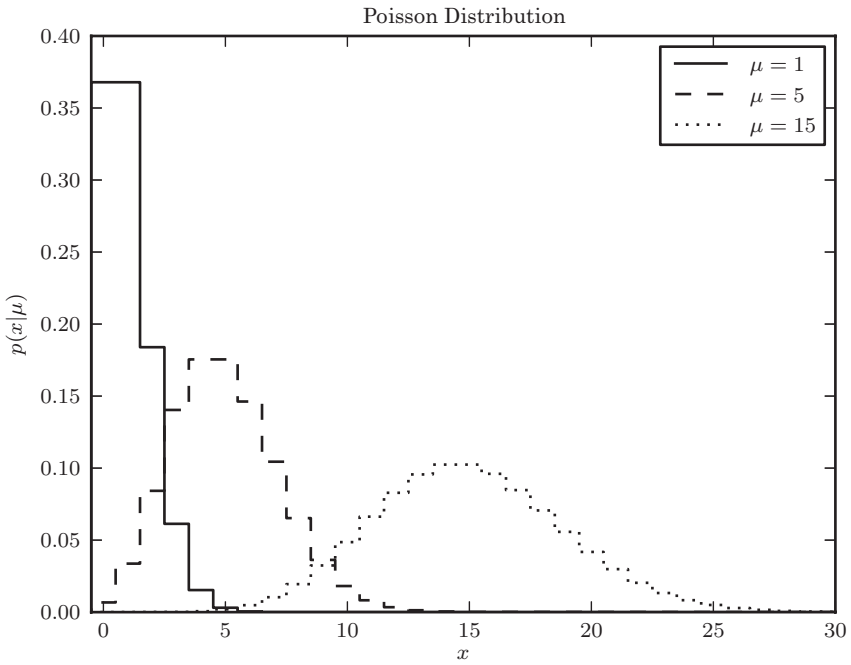

Figure 3.10. Examples of a Poisson distribution (see eq. 3.52).

The Poisson distribution is especially important in astronomy because it describes the distribution of the number of photons counted in a given interval. Even when it is replaced in practice by a Gaussian distribution for large μ, its Poissonian origin can be recognized from the relationship $\sigma^2 = \mu$ (the converse is not true).

3.3.5. The Cauchy (Lorentzian) Distribution

The Cauchy distribution is implemented in `scipy.stats.cauchy`:

```
from scipy import stats
dist = stats.cauchy(0, 1)  # mu = 0, gamma = 1
r = dist.rvs(10)  # ten random draws
p = dist.pdf(3) # pdf evaluated at x=3
```

Let us now return to distributions of a continuous variable and consider a distribution whose behavior is markedly different from any of those discussed above: the Cauchy, or Lorentzian, distribution,

$$p(x|\mu, \gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2 + (x-\mu)^2}\right). \tag{3.53}$$

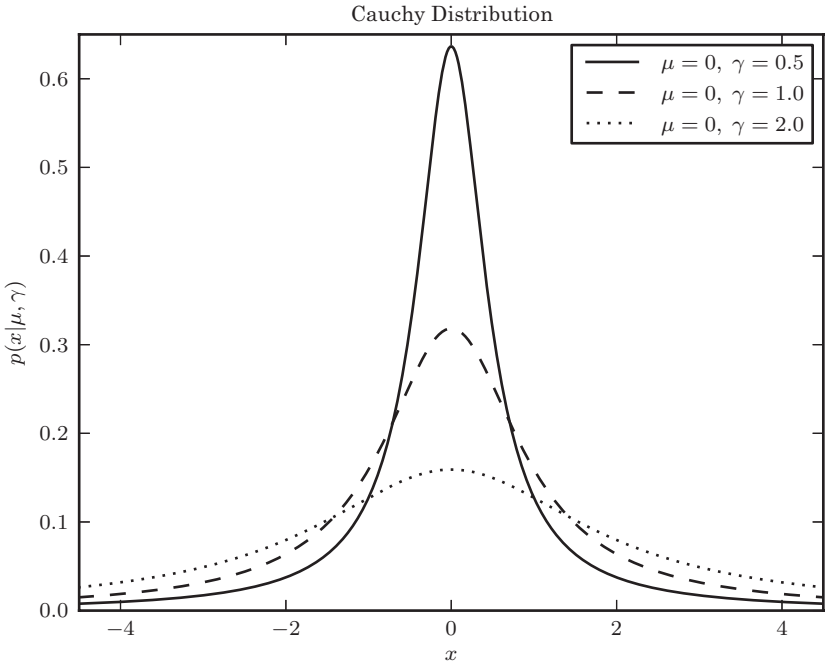

Figure 3.11. Examples of a Cauchy distribution (see eq. 3.53).

It is a symmetric distribution described by the location parameter μ and the scale parameter γ, and its median and mode are equal to μ (see figure 3.11). Because its tails decrease as slowly as x^{-2} for large $|x|$, the mean, variance, standard deviation, and higher moments do not exist. Therefore, given a set of measured x_i drawn from the Cauchy distribution, the location and scale parameters *cannot* be estimated by computing the mean value and standard deviation using standard expressions given by eqs. 3.31 and 3.32. To clarify, one can always compute a mean value for a set of numbers x_i, but this mean value will have a large scatter around μ, and furthermore, this scatter will *not* decrease with the sample size (see figure 3.12). Indeed, the mean values for many independent samples will themselves follow a Cauchy distribution. Nevertheless, μ and γ *can* be estimated as the median value and interquartile range for $\{x_i\}$. The interquartile range for the Cauchy distribution is equal to 2γ and thus

$$\sigma_G = 1.483\gamma. \tag{3.54}$$

We will see in chapter 5 how to justify this approach in a general case.

We note that the *ratio* of two independent standard normal variables ($z = (x-\mu)/\sigma$, with z drawn from $\mathcal{N}(0, 1)$) follows a Cauchy distribution with $\mu = 0$ and $\gamma = 1$. Therefore, in cases when the quantity of interest is obtained as a ratio of two other measured quantities, assuming that it is distributed as a Gaussian is a really bad idea if the quantity in the denominator has a finite chance of taking on a zero value. Furthermore, using the mean value to determine its location parameter (i.e., to get a "best" value implied by the measurements) will not achieve the "$1/\sqrt{N}$" error reduction. For a general case of the ratio of two random variables drawn from two

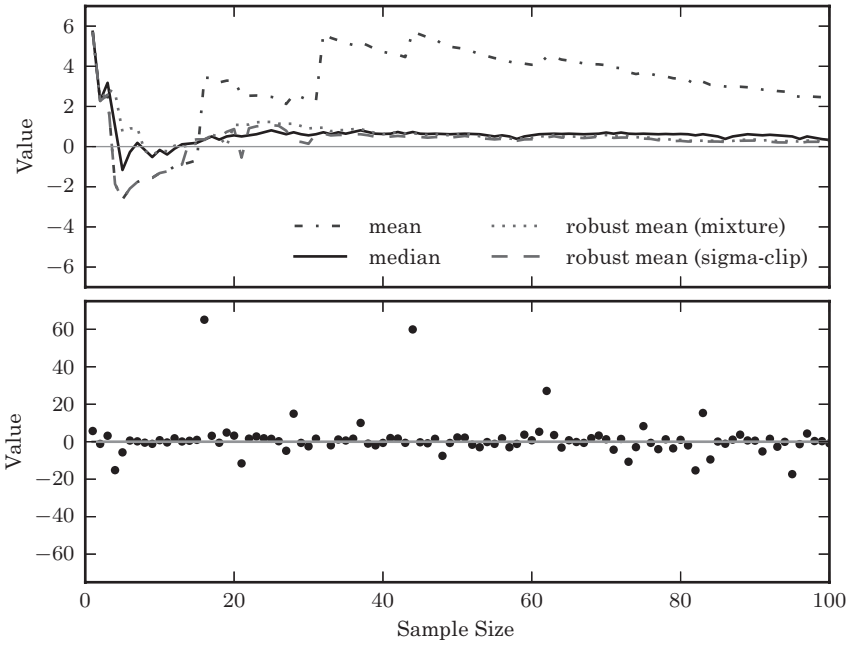

Figure 3.12. The bottom panel shows a sample of N points drawn from a Cauchy distribution with $\mu = 0$ and $\gamma = 2$. The top panel shows the sample median, sample mean, and two robust estimates of the location parameter (see text) as a function of the sample size (only points to the left from a given sample size are used). Note that the sample mean is not a good estimator of the distribution's location parameter. Though the mean appears to converge as N increases, this is deceiving: because of the large tails in the Cauchy distribution, there is always a high likelihood of a far-flung point affecting the sample mean. This behavior is markedly different from a Gaussian distribution where the probability of such "outliers" is much smaller.

different Gaussian distributions, $x = \mathcal{N}(\mu_2, \sigma_2)/\mathcal{N}(\mu_1, \sigma_1)$, the distribution of x is much more complex than the Cauchy distribution (it follows the so-called Hinkley distribution, which we will not discuss here).

Figure 3.12 also shows the results of two other robust procedures for estimating the location parameter that are often used in practice. The "clipped mean" approach computes the mean and the standard deviation for a sample (using σ_G is more robust than using eq. 3.32) and then excludes all points further than $K\sigma$ from the mean, with typically $K = 3$. This procedure is applied iteratively and typically converges very fast (in the case of a Gaussian distribution, only about 1 in 300 points is excluded). Another approach is to model the distribution as the sum of two Gaussian distributions, using methods discussed in §4.4. One of the two components is introduced to model the non-Gaussian behavior of the tails and its width is set to $K\sigma$, again with typically $K = 3$. As illustrated in figure 3.12, the location parameters estimated by these methods are similar to that of the median. In general, the performance of these methods depends on the actual distribution whose location parameter is being estimated. To summarize, when estimating the location parameter for a distribution that is not guaranteed to be Gaussian, one should use a more robust estimator than the plain sample mean given by eq. 3.31. The 3σ clipping

and the median are the simplest alternatives, but ultimately full modeling of the underlying distribution in cases when additional information is available will yield the best results.

3.3.6. The Exponential (Laplace) Distribution

The (two-sided) Laplace distribution is implemented in `scipy.stats.cauchy`, while the one-sided exponential distribution is implemented in `scipy.stats.expon`:

```
from scipy import stats
dist = stats.laplace(0, 0.5)  # mu = 0, delta = 0.5
r = dist.rvs(10)  # ten random draws
p = dist.pdf(3) # pdf evaluated at x=3
```

Similarly to a Gaussian distribution, the exponential distribution is also fully specified by only two parameters,

$$p(x|\mu, \Delta) = \frac{1}{2\Delta} \exp\left(\frac{-|x-\mu|}{\Delta}\right). \tag{3.55}$$

We note that often the exponential distribution is defined only for $x > 0$, in which case it is called the one-sided exponential distribution, and that the above expression is often called the double exponential, or Laplace distribution (see figure 3.13). The formulation adopted here is easier to use in the context of modeling error distributions than a one-sided exponential distribution. The simplest case of a one-sided exponential distribution is $p(x|\tau) = \tau^{-1} \exp(-x/\tau)$, with both the expectation (mean) value and standard deviation equal to τ. This distribution describes the time between two successive events which occur continuously and independently at a constant rate (such as photons arriving at a detector); the number of such events during some fixed time interval, T, is given by the Poisson distribution with $\mu = T/\tau$ (see eq. 3.52).

Since the Laplace distribution is symmetric around μ, its mean, mode, and median are μ, and its skewness is 0. The standard deviation is

$$\sigma = \sqrt{2}\,\Delta \approx 1.414\,\Delta, \tag{3.56}$$

and the equivalent Gaussian width estimator determined from the interquartile range ($q_{75} - q_{25} = 2(\ln 2)\Delta$) is (see eq. 3.36)

$$\sigma_G = 1.028\,\Delta. \tag{3.57}$$

Note that σ is larger than σ_G ($\sigma \approx 1.38\sigma_G$), and thus their comparison can be used to detect deviations from a Gaussian distribution toward an exponential distribution. In addition, the fractions of objects in the distribution tails are vastly

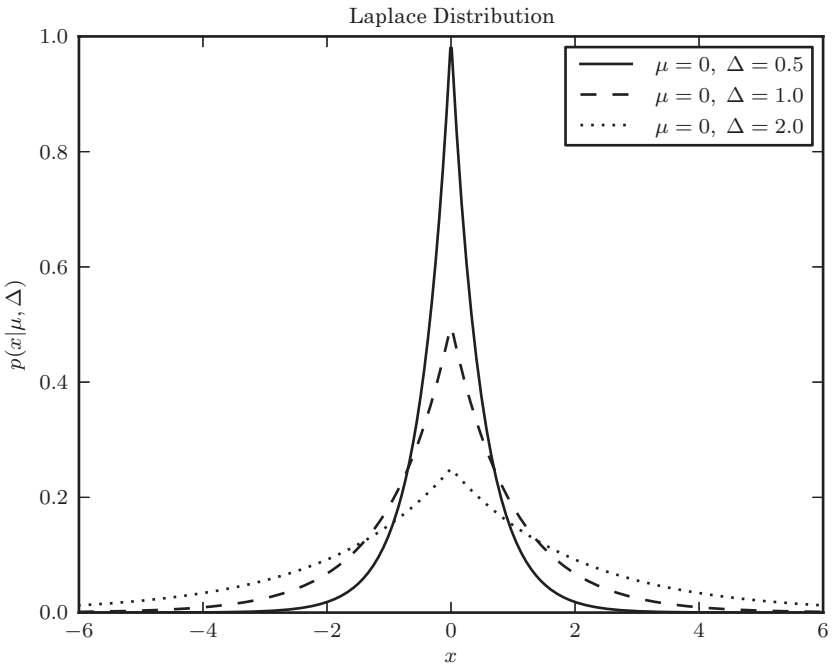

Figure 3.13. Examples of a Laplace distribution (see eq. 3.55).

different. While for the Gaussian distribution, $|x_i - \mu| > 5\sigma$ happens in fewer than one per million cases, for the exponential distribution it happens about once in a thousand cases. The kurtosis is yet another quantity that can distinguish between the Gaussian and exponential distributions; the kurtosis for the latter is 3 (see figure 3.6). Methods for comparing a measured distribution to hypothetical distributions, such as Gaussian and exponential, are discussed later in chapter 4 (§4.7) and in chapter 5.

3.3.7. The χ^2 Distribution

The χ^2 distribution is implemented in `scipy.stats.chi2`:

```
from scipy import stats
dist = stats.chi2(5)  # k = 5
r = dist.rvs(10)  # ten random draws
p = dist.pdf(1) # pdf evaluated at x=1
```

The χ^2 distribution is one of the most important distributions in statistics. If $\{x_i\}$ are drawn from a Gaussian distribution and we define $z_i = (x_i - \mu)/\sigma$, then the sum of its squares, $Q = \sum_{i=1}^{N} z_i^2$, follows a χ^2 distribution with $k = N$ degrees of freedom

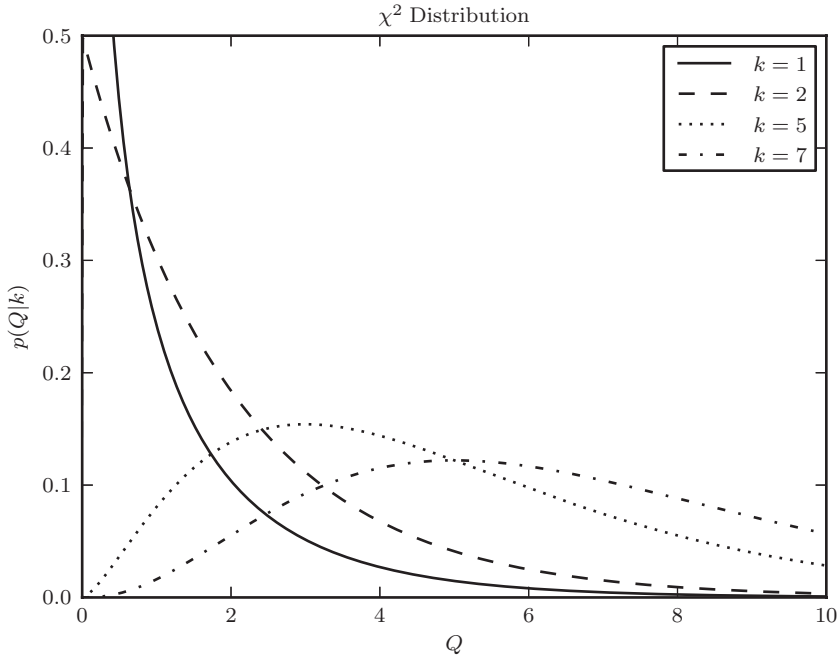

Figure 3.14. Examples of a χ^2 distribution (see eq. 3.58).

(see figure 3.14),

$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{k/2-1} \exp(-Q/2) \text{ for } Q > 0, \quad (3.58)$$

where Γ is the gamma function (for positive integers k, $\Gamma(k) \equiv (k-1)!$, and it has a closed-form expression at the half integers relevant here). In other words, the distribution of Q values depends only on the sample size, and not on the actual values of μ and σ. The importance of the χ^2 distribution will become apparent when discussing the maximum likelihood method in chapter 4.

The gamma function and log-gamma function are two of the many functions available in `scipy.special`:

```
>>> from scipy import special
>>> special.gamma(5)
24.0
>>> special.gammaln(100)  # returns log(gamma(100))
359.1342053695754
```

TABLE 3.1.
Summary of the most useful descriptive statistics for some common distributions of a continuous variable. For the definitions of listed quantities, see eqs. 3.22–3.29 and 3.36. The symbol N/A is used in cases where a simple expression does not exist, or when the quantity does not exist.

Distribution	Parameters	$\overline{x}$	q_{50}	x_m	σ	σ_G	Σ	K
Gaussian	μ, σ	μ	μ	μ	σ	σ	0	0
Uniform	μ, W	μ	μ	N/A	$W/\sqrt{12}$	$0.371W$	0	-1.2
Exponential	μ, Δ	μ	μ	μ	$\sqrt{2}\Delta$	1.028Δ	0	3
Poisson	μ	μ	$\mu - 1/3$	$\mu - 1$	$\sqrt{\mu}$	N/A	$1/\sqrt{\mu}$	$1/\mu$
Cauchy	μ, γ	N/A	μ	μ	N/A	1.483γ	N/A	N/A
$\chi^2_{\rm dof}$	k	1	$(1 - 2/9k)^3$	max $(0, 1 - 2/k)$	$\sqrt{2/k}$	N/A	$\sqrt{8/k}$	$12/k$

It is convenient to define the χ^2 distribution *per degree of freedom* as

$$\chi^2_{\rm dof}(Q|k) \equiv \chi^2(Q/k|k). \tag{3.59}$$

The mean value for $\chi^2_{\rm dof}$ is 1, the median is approximately equal to $(1 - 2/9k)^3$, and the standard deviation is $\sqrt{2/k}$. The skewness for $\chi^2_{\rm dof}$ is $\sqrt{8/k}$, and its kurtosis is $12/k$. Therefore, as k increases, $\chi^2_{\rm dof}$ tends to $\mathcal{N}(1, \sqrt{2/k})$. For example, if $k = 200$, the probability that the value of Q/k is in the 0.9–1.1 interval is 0.68. On the other hand, when $k = 20,000$, the interval containing the same probability is 0.99–1.01. The quantiles for the $\chi^2_{\rm dof}$ distribution do not have a closed-form expression.

Note that the sum Q computed for a set of x_i is very sensitive to outliers (i.e., points many σ "away" from μ). A single spurious measurement has the potential of significantly influencing analysis based on χ^2. How to deal with this problem in a general case is discussed in §4.3.

Descriptive statistics for the above distributions are summarized in table 3.1.

We review five more distributions below that frequently arise when analyzing data and comparing empirical distributions.

3.3.8. Student's *t* Distribution

Student's t distribution is implemented in `scipy.stats.t`:

```
from scipy import stats
dist = stats.t(5)  # k = 5
r = dist.rvs(10)  # ten random draws
p = dist.pdf(4) # pdf evaluated at x=4
```

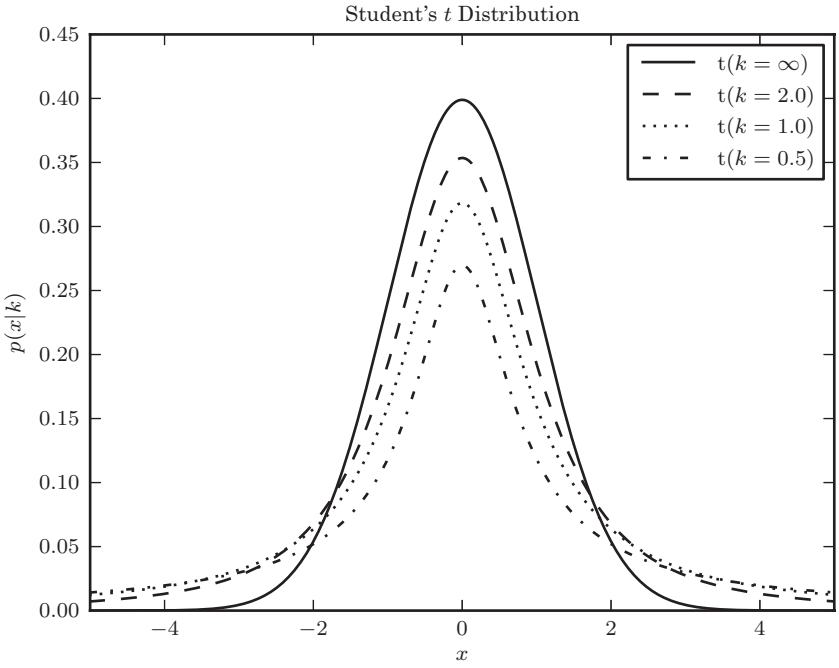

Figure 3.15. Examples of Student's t distribution (see eq. 3.60).

Student's t distribution has the probability density function

$$p(x|k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k}\,\Gamma(\frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \tag{3.60}$$

where k is the number of degrees of freedom (see figure 3.15). Note that for $k = 2$, this distribution is a Cauchy distribution with $\mu = 0$ and $\gamma = 1$ (i.e., $p(x) = \pi^{-1}(1 + x^2)^{-1}$). Student's t distribution[5] is symmetric and bell shaped, but with heavier tails than for a Gaussian distribution. The mean, median, and mode for Student's t distribution are all zero when $k > 1$ and undefined for $k = 1$. When $k > 2$, the standard deviation is $\sqrt{k/(k-2)}$; when $k > 3$, skewness is 0; and when $k > 4$, kurtosis is $6/(k-4)$. For large k, Student's t distribution tends to $\mathcal{N}(0, 1)$.

Given a sample of N measurements, $\{x_i\}$, drawn from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, the variable

$$t = \frac{\overline{x} - \mu}{s/\sqrt{N}}, \tag{3.61}$$

where $\overline{x}$ and s are given by eqs. 3.31 and 3.32, respectively, follows Student's t distribution with $k = N - 1$ degrees of freedom. Note that, although there is some

[5]It seems that just about every book on statistics must mention that Student was the pen name of W. S. Gosset, and that Gosset worked for the Guinness brewery, possibly because making beer is more fun than practicing statistics.

similarity between t defined here and Q from the χ^2 distribution, the main difference is that the definition of t is based on data-based *estimates* $\overline{x}$ and s, while the χ^2 statistic is based on *true* values μ and σ. Note that, analogously to the χ^2 distribution, we can repeatedly draw $\{x_i\}$ from Gaussian distributions with *different* μ and σ, and the corresponding t values for each sample will follow *identical* distributions as long as N stays unchanged.

The ratio of a standard normal variable and a variable drawn from the χ^2 distribution follows Student's t distribution. Although often approximated as a Gaussian distribution, the mean of a sample follows Student's t distribution (because s in the denominator is only an estimate of the true σ) and this difference may matter when samples are small. Student's t distribution also arises when comparing means of two samples. We discuss these results in more detail in chapter 5.

3.3.9. Fisher's F Distribution

Fisher's F distribution is implemented in `scipy.stats.f`:

```
from scipy import stats
dist = stats.f(2, 3)  # d1 = 2, d2 = 3
r = dist.rvs(10)  # ten random draws
p = dist.pdf(1) # pdf evaluated at x=1
```

Fisher's F distribution has the probability density function for $x \geq 0$, and parameters $d_1 > 0$ and $d_2 > 0$,

$$p(x|d_1, d_2) = C \left(1 + \frac{d_1}{d_2} x\right)^{-\frac{d_1+d_2}{2}} x^{\frac{d_1}{2}-1}, \tag{3.62}$$

where the normalization constant is

$$C = \frac{1}{B(d_1/2, d_2/2)} \left(\frac{d_1}{d_2}\right)^{d_1/2}, \tag{3.63}$$

and B is the beta function. Depending on d_1 and d_2, the appearance of the F distribution can greatly vary (see figure 3.16). When both d_1 and d_2 are large, the mean value is $d_2/(d_2-2) \approx 1$, and the standard deviation is $\sigma = \sqrt{2(d_1 + d_2)/(d_1 d_2)}$.

Fisher's F distribution describes the distribution of the ratio of two independent χ^2_{dof} variables with d_1 and d_2 degrees of freedom, and is useful when comparing the standard deviations of two samples. Also, if x_1 and x_2 are two independent random variables drawn from the Cauchy distribution with location parameter μ, then the ratio $|x_1 - \mu|/|x_2 - \mu|$ follows Fisher's F distribution with $d_1 = d_2 = 2$.

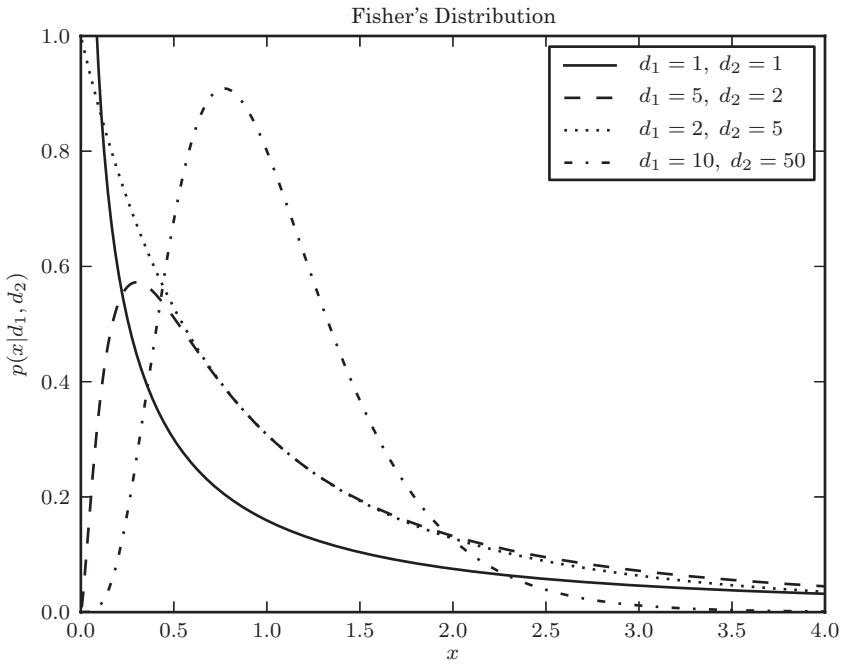

Figure 3.16. Examples of a Fisher distribution (see eq. 3.62).

3.3.10. The Beta Distribution

The beta distribution is implemented in `scipy.stats.beta`:

```
from scipy import stats
dist = stats.beta(0.5, 1.5)  # alpha = 0.5,
      # beta = 1.5
r = dist.rvs(10)  # ten random draws
p = dist.pdf(0.6) # pdf evaluated at x=0.6
```

The beta distribution is defined for $0 < x < 1$ and described by two parameters, $\alpha > 0$ and $\beta > 0$, as

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1}, \tag{3.64}$$

where Γ is the gamma function. The mean value for the beta distribution is $\alpha/(\alpha+\beta)$. Various combinations of the parameters α and β can result in very different shapes (see figure 3.17). This property makes it very useful in Bayesian analysis, as discussed in §5.6.2. In particular, the beta distribution is the conjugate prior for the binomial distribution (see §5.2.3).

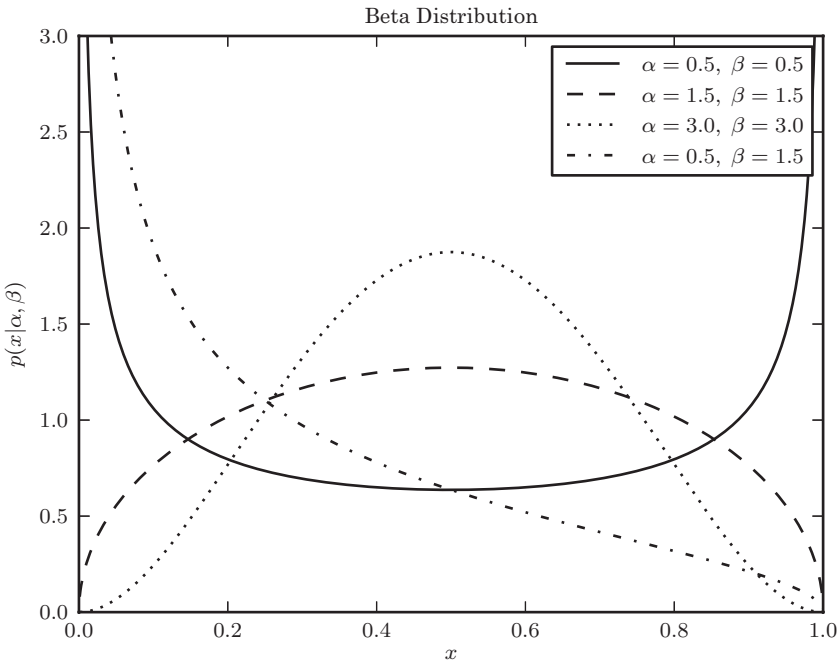

Figure 3.17. Examples of the beta distribution (see eq. 3.64).

3.3.11. The Gamma Distribution

The gamma distribution is implemented in `scipy.stats.gamma`:

```
from scipy import stats
dist = stats.gamma(1, 0, 2)  # k = 1, loc = 0,
       # theta = 2
r = dist.rvs(10)  # ten random draws
p = dist.pdf(1) # pdf evaluated at x=1
```

The gamma distribution is defined for $0 < x < \infty$, and is described by two parameters, a shape parameter k and a scale parameter θ (see figure 3.18). The probability distribution function is given by

$$p(x|k, \theta) = \frac{1}{\theta^k} \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)}, \tag{3.65}$$

where $\Gamma(k)$ is the gamma function.

The gamma distribution is useful in Bayesian statistics, as it is a conjugate prior to several distributions including the exponential (Laplace) distribution, and the Poisson distribution (see §5.2.3).

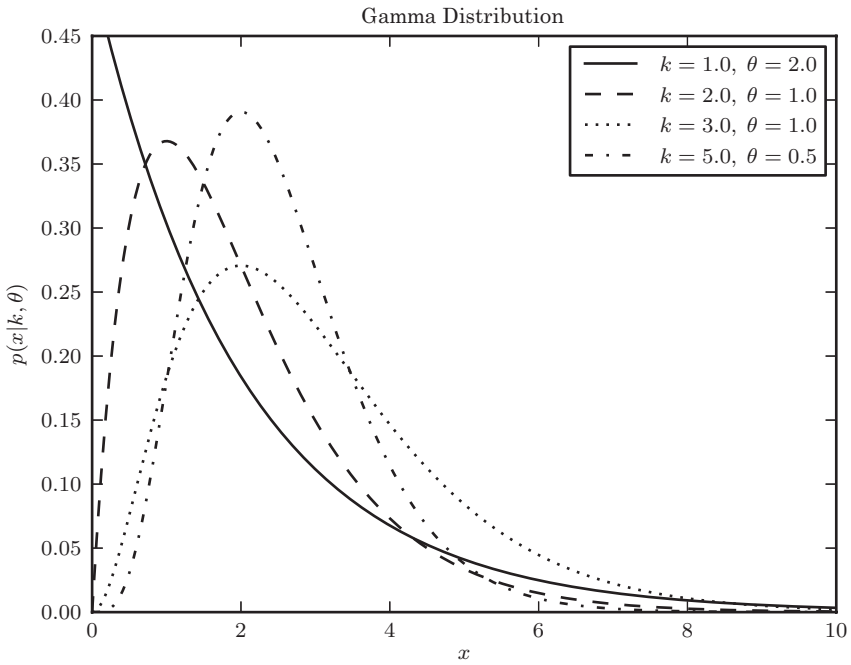

Figure 3.18. Examples of a gamma distribution (see eq. 3.65).

3.3.12. The Weibull Distribution

The Weibull distribution is implemented in `scipy.stats.dweibull`:

```
from scipy import stats
dist = stats.dweibull(1, 0, 2)  # k = 1, loc = 0,
       # lambda = 2
r = dist.rvs(10)  # ten random draws
p = dist.pdf(1) # pdf evaluated at x=1
```

The Weibull distribution is defined for $x \geq 0$ and described by the shape parameter k and the scale parameter λ as

$$p(x|k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \tag{3.66}$$

(see figure 3.19). The mean value is $\lambda\,\Gamma(1 + 1/k)$ and the median is $\lambda\,(\ln 2)^{1/k}$. The shape parameter k can be used to smoothly interpolate between the exponential distribution (corresponding to $k = 1$; see §3.3.6) and the Rayleigh distribution ($k = 2$; see §3.3.7). As k tends to infinity, the Weibull distribution morphs into a Dirac δ function. If x is uniformly distributed on the interval (0, 1), then the random

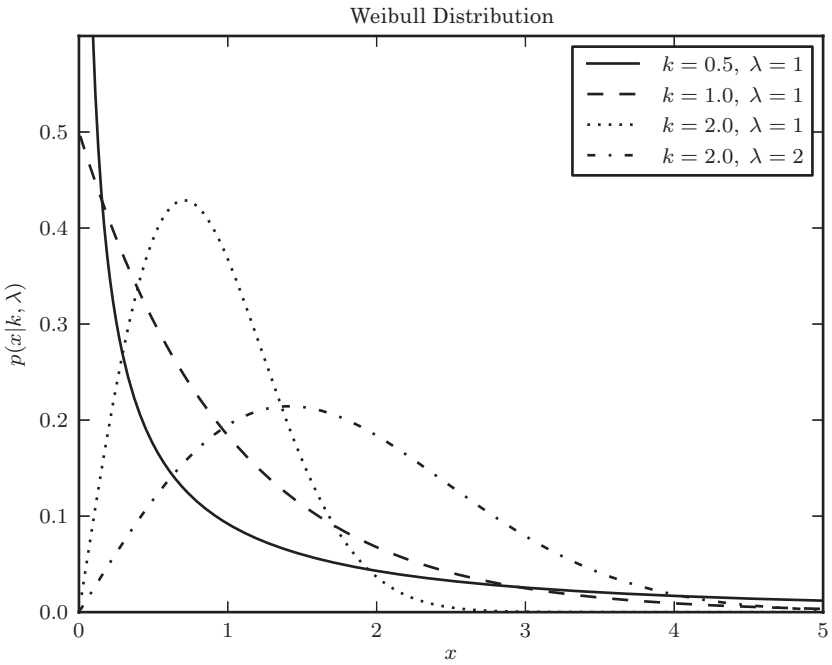

Figure 3.19. Examples of a Weibull distribution (see eq. 3.66).

variable $y = \lambda(-\ln x)^{1/k}$ is distributed as the Weibull distribution with parameters k and λ.

The Weibull distribution is often encountered in physics and engineering because it provides a good description of a random failure process with a variable rate, wind behavior, distribution of extreme values, and the size distribution of particles. For example, if x is the time to failure for a device with the failure rate proportional to a power of time, t^m, then x follows the Weibull distribution with $k = m + 1$. The distribution of the wind speed at a given location typically follows the Weibull distribution with $k \approx 2$ (the $k = 2$ case arises when two components of a two-dimensional vector are uncorrelated and follow Gaussian distributions with equal variances).

In an engineering context, an elegant method was developed to assess whether data $\{x_i\}$ was drawn from the Weibull distribution. The method utilizes the fact that the cumulative distribution function for the Weibull distribution is very simple:

$$H_W(x) = 1 - e^{-(x/\lambda)^k}. \tag{3.67}$$

The cumulative distribution function based on data, $F(x)$, is used to define $z = \ln(-\ln(1 - F(x)))$ and z is plotted as a function of $\ln x$. For the Weibull distribution, the z vs. $\ln x$ plot is a straight line with a slope equal to k and intercept equal to $-k \ln \lambda$. For small samples, it is difficult to distinguish the Weibull and log-normal distributions.

3.4. The Central Limit Theorem

The central limit theorem provides the theoretical foundation for the practice of repeated measurements in order to improve the accuracy of the final result. Given an *arbitrary* distribution $h(x)$, characterized by its mean μ and standard deviation σ, the central limit theorem says that the mean of N values x_i drawn from that distribution will approximately follow a Gaussian distribution $\mathcal{N}(\mu, \sigma/\sqrt{N})$, with the approximation accuracy improving with N. This is a remarkable result since the details of the distribution $h(x)$ are not specified—we can "average" our measurements (i.e., compute their mean value using eq. 3.31) and expect the $1/\sqrt{N}$ improvement in accuracy *regardless of details in our measuring apparatus!* The underlying reason why the central limit theorem can make such a far-reaching statement is the strong assumption about $h(x)$: it must have a standard deviation and thus its tails must fall off faster than $1/x^2$ for large x. As more measurements are combined, the tails will be "clipped" and eventually (for large N) the mean will follow a Gaussian distribution (it is easy to prove this theorem using standard tools from statistics such as characteristic functions; e.g., see Lup93). Alternatively, it can be shown that the resulting Gaussian distribution rises as the result of many consecutive convolutions (e.g., see Greg05). An illustration of the central limit theorem in action, using a uniform distribution for $h(x)$, is shown in figure 3.20.

However, there are cases when the central limit theorem *cannot* be invoked! We already discussed the Cauchy distribution, which does not have a well-defined mean or standard deviation, and thus the central limit theorem is *not applicable* (recall figure 3.12). In other words, if we repeatedly draw N values x_i from a Cauchy distribution and compute their mean value, the resulting distribution of these mean values will *not* follow a Gaussian distribution (it will follow the Cauchy distribution, and will have an infinite variance). If we decide to use the mean of measured values to estimate the location parameter μ, we will *not* gain the $\sqrt{N}$ improvement in accuracy promised by the central limit theorem. Instead, we need to compute the median and interquartile range for x_i, which are unbiased estimators of the location and scale parameters for the Cauchy distribution. Of course, the reason why the central limit theorem is not applicable to the Cauchy distribution is its extended tails that decrease only as x^{-2}.

We mention in passing the weak law of large numbers (also known as Bernoulli's theorem): the sample mean converges to the distribution mean as the sample size increases. Again, for distributions with ill-defined variance, such as the Cauchy distribution, *the weak law of large numbers breaks down.*

In another extreme case of tail behavior, we have the uniform distribution which does not even have tails (cf. §3.3.1). If we repeatedly draw N values x_i from a uniform distribution described by its mean μ and width W, the distribution of their mean value $\overline{x}$ will be centered on μ, as expected from the central limit theorem. In addition, the uncertainty of our estimate for the location parameter μ will decrease proportionally to $1/\sqrt{N}$, again in agreement with the central limit theorem. However, using the mean to estimate μ is *not* the best option here, and indeed μ *can be estimated with an accuracy that improves as* $1/N$*, that is, faster than* $1/\sqrt{N}$.

How is this arguably surprising result possible? Given the uniform distribution described by eq. 3.39, a value x_i that happens to be larger than μ rules out all

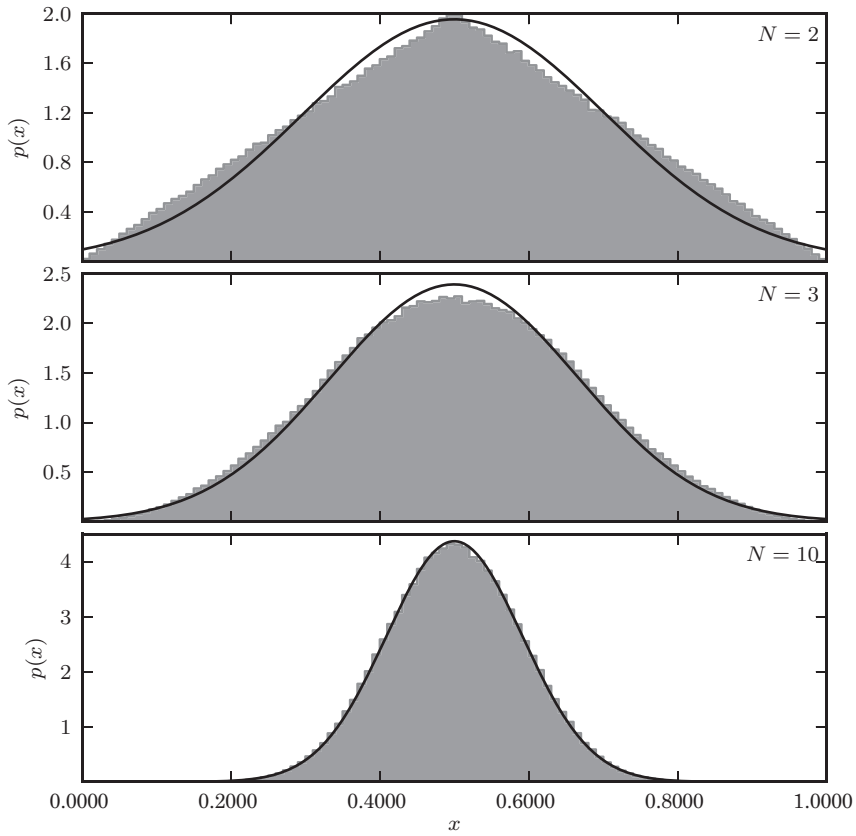

Figure 3.20. An illustration of the central limit theorem. The histogram in each panel shows the distribution of the mean value of N random variables drawn from the $(0, 1)$ range (a uniform distribution with $\mu = 0.5$ and $W = 1$; see eq. 3.39). The distribution for $N = 2$ has a triangular shape and as N increases it becomes increasingly similar to a Gaussian, in agreement with the central limit theorem. The predicted normal distribution with $\mu = 0.5$ and $\sigma = 1/\sqrt{12N}$ is shown by the line. Already for $N = 10$, the "observed" distribution is essentially the same as the predicted distribution.

values $\mu < x_i - W/2$. This strong conclusion is of course the result of the sharp edges of the uniform distribution. The strongest constraint on μ comes from the extremal value of x_i and thus we know that $\mu > \max(x_i) - W/2$. Analogously, we know that $\mu < \min(x_i) + W/2$ (of course, it must be true that $\max(x_i) \leq W/2$ and $\min(x_i) \geq -W/2$). Therefore, given N values x_i, the allowed range for μ is $\max(x_i) - W/2 < \mu < \min(x_i) + W/2$, with a uniform probability distribution for μ within that range. The best estimate for μ is then in the middle of the range,

$$\tilde{\mu} = \frac{\min(x_i) + \max(x_i)}{2}, \tag{3.68}$$

and the standard deviation of this estimate (note that the scatter of this estimate around the true value μ is not Gaussian) is the width of the allowed interval, R,

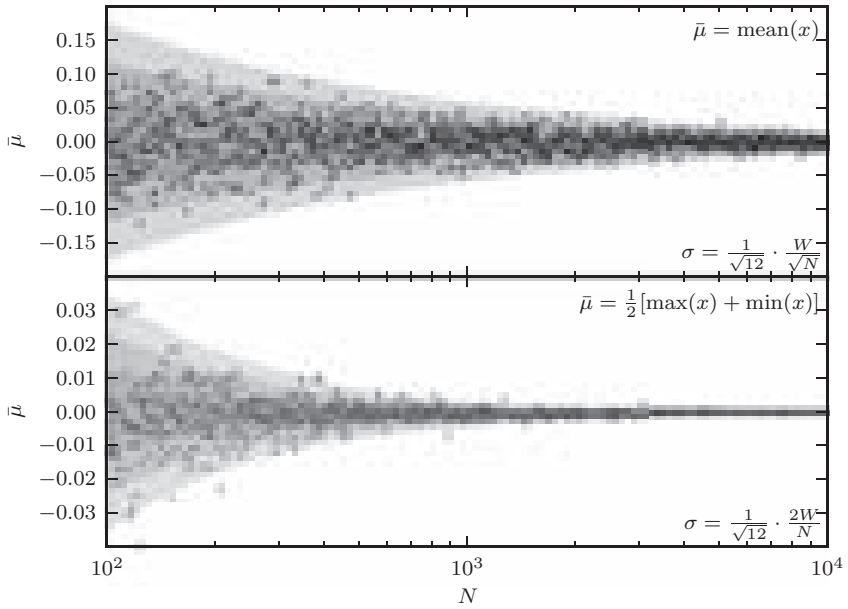

Figure 3.21. A comparison of the sample-size dependence of two estimators for the location parameter of a uniform distribution, with the sample size ranging from $N = 100$ to $N = 10{,}000$. The estimator in the top panel is the sample mean, and the estimator in the bottom panel is the mean value of two extreme values. The theoretical 1σ, 2σ, and 3σ contours are shown for comparison. When using the sample mean to estimate the location parameter, the uncertainty decreases proportionally to $1/\sqrt{N}$, and when using the mean of two extreme values as $1/N$. Note different vertical scales for the two panels.

divided by $\sqrt{12}$ (cf. eq. 3.40). In addition, the best estimate for W is given by

$$\tilde{W} = [\max(x_i) - \min(x_i)] \frac{N}{N-2}. \tag{3.69}$$

What is the width of the allowed interval, $R = (\max(x_i) - \min(x_i) - W)$? By considering the distribution of extreme values of x_i, it can be shown that the expectation values are $E[\min(x_i)] = (\mu - W/2 + W/N)$ and $E[\max(x_i)] = (\mu + W/2 - W/N)$. These results can be easily understood: if N values x_i are uniformly scattered within a box of width W, then the two extreme points will be on average $\sim W/N$ away from the box edges. Therefore, the width of the allowed range for μ is $R = 2W/N$, and $\tilde{\mu}$ is an unbiased estimator of μ with a standard deviation of

$$\sigma_{\tilde{\mu}} = \frac{2W}{\sqrt{12}\,N}. \tag{3.70}$$

While the mean value of x_i is also an unbiased estimator of μ, $\tilde{\mu}$ is a much more *efficient* estimator: the ratio of the two uncertainties is $2/\sqrt{N}$ and $\tilde{\mu}$ wins for $N > 2$. The different behavior of these two estimators is illustrated in figure 3.21.

In summary, while the central limit theorem is of course valid for the uniform distribution, the mean of x_i is not the most efficient estimator of the location

parameter μ. Due to the absence of tails, the distribution of extreme values of x_i provides the most efficient estimator, $\tilde{\mu}$, which improves with the sample size as fast as $1/N$.

The Cauchy distribution and the uniform distribution are vivid examples of cases where taking the mean of measured values is not an appropriate procedure for estimating the location parameter. What do we do in a general case when the optimal procedure is not known? We will see in chapter 5 that maximum likelihood and Bayesian methods offer an elegant general answer to this question (see §5.6.4).

3.5. Bivariate and Multivariate Distribution Functions

3.5.1. Two-Dimensional (Bivariate) Distributions

All the distribution functions discussed so far are one-dimensional: they describe the distribution of N measured values x_i. Let us now consider the case when two values are measured in each instance: x_i and y_i. Let us assume that they are drawn from a two-dimensional distribution described by $h(x, y)$, with $\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} h(x, y)\, dy = 1$. The distribution $h(x, y)$ should be interpreted as giving the probability that x is between x and $x + dx$ *and* that y is between y and $y + dy$.

In analogy with eq. 3.23, the two variances are defined as

$$V_x = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2\, h(x, y)\, dx\, dy \tag{3.71}$$

and

$$V_y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \mu_y)^2\, h(x, y)\, dx\, dy, \tag{3.72}$$

where the mean values are defined as

$$\mu_x = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, h(x, y)\, dx\, dy \tag{3.73}$$

and analogously for μ_y. In addition, the covariance of x and y, which is a measure of the dependence of the two variables on each other, is defined as

$$V_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y)\, h(x, y)\, dx\, dy. \tag{3.74}$$

Sometimes, Cov(x,y) is used instead of V_{xy}. For later convenience, we define $\sigma_x = \sqrt{V_x}$, $\sigma_y = \sqrt{V_y}$, and $\sigma_{xy} = V_{xy}$ (note that there is no square root; i.e., the unit for σ_{xy} is the square of the unit for σ_x and σ_y). A very useful related result is that the variance of the sum $z = x + y$ is

$$V_z = V_x + V_y + 2\, V_{xy}. \tag{3.75}$$

When x and y are uncorrelated ($V_{xy} = 0$), the variance of their sum is equal to the sum of their variances. For $w = x - y$,

$$V_w = V_x + V_y - 2\,V_{xy}. \tag{3.76}$$

In the two-dimensional case, it is important to distinguish the marginal distribution of one variable, for example, here for x:

$$m(x) = \int_{-\infty}^{\infty} h(x, y)\,dy, \tag{3.77}$$

from the two-dimensional distribution evaluated at a given $y = y_o$, $h(x, y_o)$ (and analogously for y). The former is generally wider than the latter, as will be illustrated below using a Gaussian example. Furthermore, while $m(x)$ is a properly normalized probability distribution ($\int_{-\infty}^{\infty} m(x)\,dx = 1$), $h(x, y = y_o)$ is not (recall the discussion in §3.1.3).

If $\sigma_{xy} = 0$, then x and y are uncorrelated and we can treat them separately as two independent one-dimensional distributions. Here "independence" means that whatever range we impose on one of the two variables, the distribution of the other one remains unchanged. More formally, we can describe the underlying two-dimensional probability distribution function as the product of two functions that each depends on only one variable:

$$h(x, y) = h_x(x)\,h_y(y). \tag{3.78}$$

Note that in this special case, marginal distributions are identical to h_x and h_y, and $p(x|y = y_o)$ is the same as $h_x(x)$ except for different normalization.

3.5.2. Bivariate Gaussian Distributions

A generalization of the Gaussian distribution to the two-dimensional case is given by

$$p(x, y|\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(\frac{-z^2}{2(1-\rho^2)}\right), \tag{3.79}$$

where

$$z^2 = \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - 2\,\rho\,\frac{(x-\mu_x)\,(y-\mu_y)}{\sigma_x\,\sigma_y}, \tag{3.80}$$

and the (dimensionless) correlation coefficient between x and y is defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{3.81}$$

(see figure 3.22). For perfectly correlated variables such that $y = ax + b$, $\rho = a/|a| \equiv \text{sign}(a)$, and for uncorrelated variables, $\rho = 0$. The *population* correlation coefficient ρ is directly related to Pearson's *sample* correlation coefficient r discussed in §3.6.

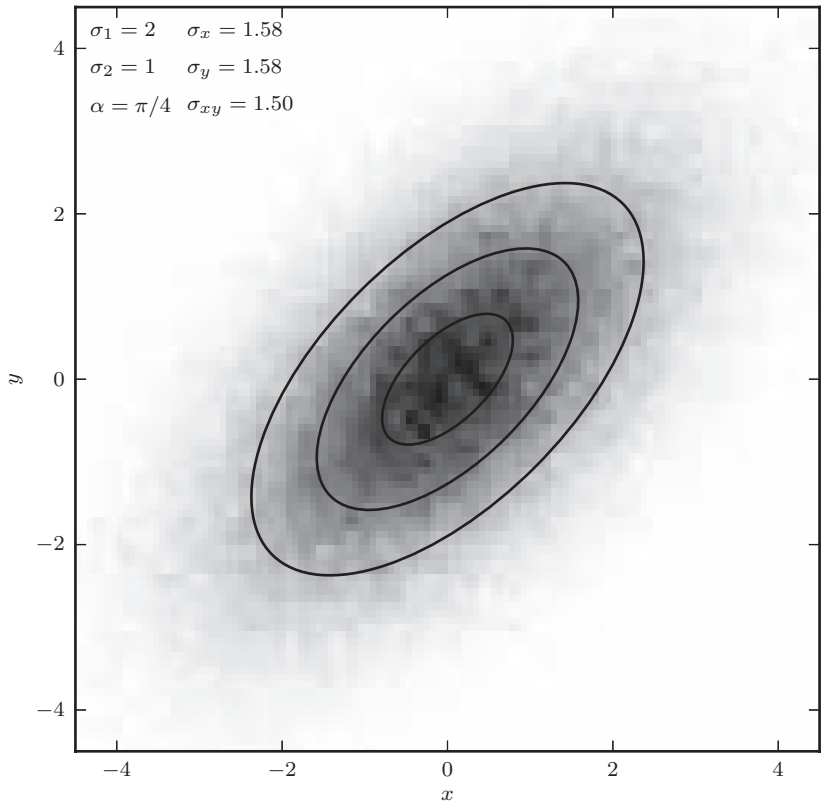

Figure 3.22. An example of data generated from a bivariate Gaussian distribution. The shaded pixels are a Hess diagram showing the density of points at each position.

The contours in the (x, y) plane defined by $p(x, y|\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}) =$ constant are ellipses centered on $(x = \mu_x, y = \mu_y)$, and the angle α (defined for $-\pi/2 \leq \alpha \leq \pi/2$) between the x-axis and the ellipses' major axis is given by

$$\tan(2\alpha) = 2\,\rho\,\frac{\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} = 2\frac{\sigma_{xy}}{\sigma_x^2 - \sigma_y^2}. \tag{3.82}$$

When the (x, y) coordinate system is rotated by an angle α around the point $(x = \mu_x, y = \mu_y)$,

$$\begin{aligned} P_1 &= (x - \mu_x)\cos\alpha + (y - \mu_y)\sin\alpha, \\ P_2 &= -(x - \mu_x)\sin\alpha + (y - \mu_y)\cos\alpha, \end{aligned} \tag{3.83}$$

the correlation between the two new variables P_1 and P_2 disappears, and the two widths are

$$\sigma_{1,2}^2 = \frac{\sigma_x^2 + \sigma_y^2}{2} \pm \sqrt{\left(\frac{\sigma_x^2 - \sigma_y^2}{2}\right)^2 + \sigma_{xy}^2}. \tag{3.84}$$

The coordinate axes P_1 and P_2 are called the principal axes, and σ_1 and σ_2 represent the minimum and maximum widths obtainable for any rotation of the coordinate axes. In this coordinate system where the correlation vanishes, the bivariate Gaussian is the product of two univariate Gaussians (see eq. 3.78). We shall discuss a multidimensional extension of this idea (principal component analysis) in chapter 7.

Alternatively, starting from the principal axes frame, we can compute

$$\sigma_x = \sqrt{\sigma_1^2 \cos^2\alpha + \sigma_2^2 \sin^2\alpha}, \tag{3.85}$$

$$\sigma_y = \sqrt{\sigma_1^2 \sin^2\alpha + \sigma_2^2 \cos^2\alpha}, \tag{3.86}$$

and ($\sigma_1 \geq \sigma_2$ by definition)

$$\sigma_{xy} = (\sigma_1^2 - \sigma_2^2) \sin\alpha \cos\alpha. \tag{3.87}$$

Note that σ_{xy}, and thus the correlation coefficient ρ, vanish for both $\alpha = 0$ and $\alpha = \pi/2$, and have maximum values for $\pi/4$. By inverting eq. 3.83, we get

$$\begin{aligned} x &= \mu_x + P_1 \cos\alpha - P_2 \sin\alpha, \\ y &= \mu_y + P_1 \sin\alpha + P_2 \cos\alpha. \end{aligned} \tag{3.88}$$

These expressions are very useful when generating mock samples based on bivariate Gaussians (see §3.7).

The marginal distribution of the y variable is given by

$$m(y|I) = \int_{-\infty}^{\infty} p(x, y|I)\, dx = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(\frac{-(y-\mu_y)^2}{2\sigma_y^2}\right), \tag{3.89}$$

where we used shorthand $I = (\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy})$, and analogously for $m(x)$. Note that $m(y|I)$ does not depend on μ_x, σ_x, and σ_{xy}, and it is equal to $\mathcal{N}(\mu_y, \sigma_y)$. Let us compare $m(y|I)$ to $p(x, y|I)$ evaluated for the most probable x,

$$p(x = \mu_x, y|I) = \frac{1}{\sigma_x\sqrt{2\pi}} \frac{1}{\sigma_*\sqrt{2\pi}} \exp\left(\frac{-(y-\mu_y)^2}{2\sigma_*^2}\right) = \frac{1}{\sigma_x\sqrt{2\pi}} \mathcal{N}(\mu_y, \sigma_*), \tag{3.90}$$

where

$$\sigma_* = \sigma_y\sqrt{1-\rho^2} \leq \sigma_y. \tag{3.91}$$

Since $\sigma_* \leq \sigma_y$, $p(x = \mu_x, y|I)$ is narrower than $m(y|I)$, reflecting the fact that the latter carries additional uncertainty due to unknown (marginalized) x. It is generally true that $p(x, y|I)$ evaluated for any fixed value of x will be proportional to a Gaussian with the width equal to σ_* (and centered on the P_1-axis). In other words,

eq. 3.79 can be used to "predict" the value of y for an arbitrary x when $\mu_x, \mu_y, \sigma_x, \sigma_y$, and σ_{xy} are estimated from a given data set.

In the next section we discuss how to estimate the parameters of a bivariate Gaussian $(\mu_x, \mu_y, \sigma_1, \sigma_2, \alpha)$ using a set of points (x_i, y_i) whose uncertainties are negligible compared to σ_1 and σ_2. We shall return to this topic when discussing regression methods in chapter 8, including the fitting of linear models to a set of points (x_i, y_i) whose measurement uncertainties (i.e., *not* their distribution) are described by an analog of eq. 3.79.

3.5.3. A Robust Estimate of a Bivariate Gaussian Distribution from Data

AstroML provides a routine for both the robust and nonrobust estimates of the parameters for a bivariate normal distribution:

```
# assume x and y are pre-defined data arrays
from astroML.stats import fit_bivariate_normal
mean, sigma1, sigma2, alpha =\
        fit_bivariate_
    normal(x, y)
```

For further examples, see the source code associated with figure 3.23.

A bivariate Gaussian distribution is often encountered in practice when dealing with two-dimensional problems, and typically we need to estimate its parameters using data vectors x and y. Analogously to the one-dimensional case, where we can estimate parameters μ and σ as $\overline{x}$ and s using eqs. 3.31 and 3.32, here we can estimate the five parameters $(\overline{x}, \overline{y}, s_x, s_y, s_{xy})$ using similar equations that correspond to eqs. 3.71–3.74. In particular, the correlation coefficient is estimated using Pearson's sample correlation coefficient, r (eq. 3.102, discussed in §3.6). The principal axes can be easily found with α estimated using

$$\tan(2\alpha) = 2\frac{s_x s_y}{s_x^2 - s_y^2} r, \tag{3.92}$$

where for simplicity we use the same symbol for both population and sample values of α.

When working with real data sets that often have outliers (i.e., a small fraction of points are drawn from a significantly different distribution than for the majority of the sample), eq. 3.92 can result in grossly incorrect values of α because of the impact of outliers on s_x, s_y, and r. A good example is the measurement of the velocity ellipsoid for a given population of stars, when another population with vastly different kinematics contaminates the sample (e.g., halo vs. disk stars). A simple and efficient remedy is to use the median instead of the mean, and to use the interquartile range to estimate variances.

While it is straightforward to estimate s_x and s_y from the interquartile range (see eq. 3.36), it is not so for s_{xy}, or equivalently, r. To robustly estimate r, we can use the

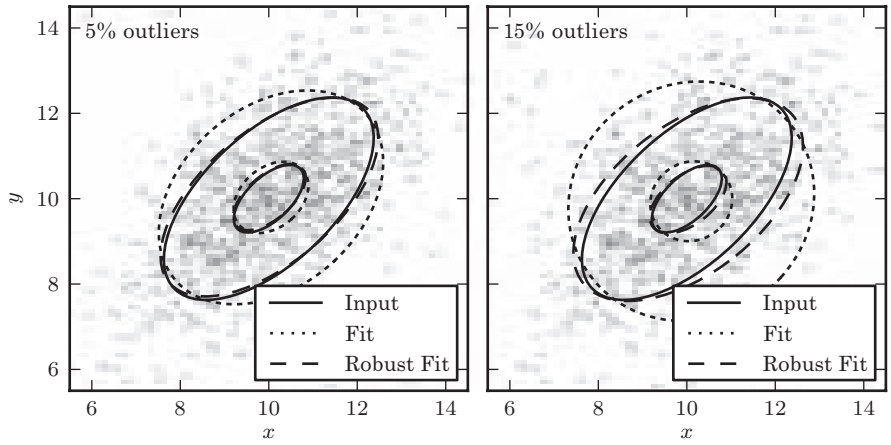

Figure 3.23. An example of computing the components of a bivariate Gaussian using a sample with 1000 data values (points), with two levels of contamination. The core of the distribution is a bivariate Gaussian with $(\mu_x, \mu_y, \sigma_1, \sigma_2, \alpha) = (10, 10, 2, 1, 45^\circ)$. The "contaminating" subsample contributes 5% (left) and 15% (right) of points centered on the same (μ_x, μ_y), and with $\sigma_1 = \sigma_2 = 5$. Ellipses show the 1σ and 3σ contours. The solid lines correspond to the input distribution. The thin dotted lines show the nonrobust estimate, and the dashed lines show the robust estimate of the best-fit distribution parameters (see §3.5.3 for details).

following identity for the correlation coefficient (for details and references, see [5]):

$$\rho = \frac{V_u - V_w}{V_u + V_w}, \tag{3.93}$$

where V stands for variance, and transformed coordinates are defined as $(\mathrm{Cov}(u,w) = 0)$

$$u = \frac{\sqrt{2}}{2}\left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y}\right) \tag{3.94}$$

and

$$w = \frac{\sqrt{2}}{2}\left(\frac{x}{\sigma_x} - \frac{y}{\sigma_y}\right). \tag{3.95}$$

By substituting the robust estimator σ_G^2 in place of the variance V in eq. 3.93, we can compute a robust estimate of r, and in turn a robust estimate of the principal axis angle α. Error estimates for r and α can be easily obtained using the bootstrap and jackknife methods discussed in §4.5. Figure 3.23 illustrates how this approach helps when the sample is contaminated by outliers. For example, when the fraction of contaminating outliers is 15%, the best-fit α determined using the nonrobust method is grossly incorrect, while the robust best fit still recognizes the orientation of the distribution's core. Even when outliers contribute only 5% of the sample, the robust estimate of σ_2/σ_1 is much closer to the input value.

3.5.4. Multivariate Gaussian Distributions

The function `multivariate_normal` in the module `numpy.random` implements random samples from a multivariate Gaussian distribution:

```
>>> import numpy as np
>>> mu = [1, 2]
>>> cov = [[1, 0.2],
...        [0.2, 3]]
>>> np.random.multivariate_normal(mu, cov)
array([ 0.03438156, -2.60831303])
```

This was a two-dimensional example, but the function can handle any number of dimensions.

Analogously to the two-dimensional (bivariate) distribution given by eq. 3.79, the Gaussian distribution can be extended to multivariate Gaussian distributions in an arbitrary number of dimensions. Instead of introducing new variables by name, as we did by adding y to x in the bivariate case, we introduce a vector variable $\mathbf{x}$ (i.e., instead of a scalar variable x). We use M for the problem dimensionality ($M = 2$ for the bivariate case) and thus the vector $\mathbf{x}$ has M components. In the one-dimensional case, the variable x has N values x_i. In the multivariate case, each of M components of $\mathbf{x}$, let us call them x^j, $j = 1, \ldots, M$, has N values denoted by x_i^j. With the aid of linear algebra, results from the preceding section can be expressed in terms of matrices, and then trivially extended to an arbitrary number of dimensions. The notation introduced here will be extensively used in later chapters.

The argument of the exponential function in eq. 3.79 can be rewritten as

$$\arg = -\frac{1}{2}\left(\alpha x^2 + \beta y^2 + 2\gamma xy\right), \tag{3.96}$$

with σ_x, σ_y, and σ_{xy} expressed as functions of α, β, and γ (e.g., $\sigma_x^2 = \beta/(\alpha\beta - \gamma^2)$), and the distribution is centered on the origin for simplicity (we could replace $\mathbf{x}$ by $\mathbf{x} - \overline{\mathbf{x}}$, where $\overline{\mathbf{x}}$ is the vector of mean values, if need be). This form lends itself better to matrix notation:

$$p(\mathbf{x}|I) = \frac{1}{(2\pi)^{M/2}\sqrt{\det(\mathbf{C})}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}\right), \tag{3.97}$$

where $\mathbf{x}$ is a column vector, $\mathbf{x}^T$ is its transposed row vector, $\mathbf{C}$ is the covariance matrix and $\mathbf{H}$ is equal to the inverse of the covariance matrix, $\mathbf{C}^{-1}$ (note that $\mathbf{H}$ is a symmetric matrix with positive eigenvalues).

Analogously to eq. 3.74, the elements of the covariance matrix $\mathbf{C}$ are given by

$$C_{kj} = \int_{-\infty}^{\infty} x^k x^j p(\mathbf{x}|I)\, d^M x. \tag{3.98}$$

The argument of the exponential function can be expanded in component form as

$$\mathbf{x}^T\mathbf{H}\mathbf{x} = \sum_{k=1}^{M}\sum_{j=1}^{M} H_{kj}x^k x^j. \tag{3.99}$$

For example, in the bivariate case discussed above we had $\mathbf{x}^T = (x, y)$ and

$$H_{11} = \alpha,\ H_{22} = \beta,\ H_{12} = H_{21} = \gamma, \tag{3.100}$$

with

$$\det(\mathbf{C}) = \sigma_x^2\sigma_y^2 - \sigma_{xy}^2. \tag{3.101}$$

Equivalently, starting with $C_{11} = \sigma_x^2$, $C_{12} = C_{21} = \sigma_{xy}$, and $C_{22} = \sigma_y^2$, and using $\mathbf{H} = \mathbf{C}^{-1}$ and eq. 3.97, it is straightforward to recover eq. 3.79.

Similarly to eq. 3.89, when multivariate $p(\mathbf{x}|I)$ is marginalized over all but one dimension, the result is a univariate (one-dimensional) Gaussian distribution.

3.6. Correlation Coefficients

Several correlation tests are implemented in `scipy.stats`, including `spearmanr`, `kendalltau`, and `pearsonr`:

```
from scipy import stats
x, y = np.random.random((2, 100))  # two random
       # arrays
corr_coeff, p_value = stats.pearsonr(x, y)
rho, p_value = stats.spearmanr(x, y)
tau, p_value = stats.kendalltau(x, y)
```

We have already introduced the covariance of x and y (see eq. 3.74) and the correlation coefficient (see eq. 3.81) as measures of the dependence of the two variables on each other. Here we extend our discussion to the interpretation of the sample correlation coefficient.

Given two data sets of equal size N, $\{x_i\}$ and $\{y_i\}$, Pearson's sample correlation coefficient is

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}, \tag{3.102}$$

with $-1 \le r \le 1$. For uncorrelated variables, $r = 0$. If the pairs (x_i, y_i) are drawn from two uncorrelated univariate Gaussian distributions (i.e., the population

correlation coefficient $\rho = 0$), then the distribution of r follows Student's t distribution with $k = N - 2$ degrees of freedom and

$$t = r\sqrt{\frac{N-2}{1-r^2}}. \tag{3.103}$$

Given this known distribution, a measured value of r can be transformed into the significance of the statement that $\{x_i\}$ and $\{y_i\}$ are correlated. For example, if N=10, the probability that a value of r as large as 0.72 would arise by chance is 1% (the one-sided 99% confidence level for Student's t distribution with $k = 8$ degrees of freedom is $t = 2.896$). We shall return to such an analysis in §4.6 on hypothesis testing.

When the sample is drawn from a bivariate Gaussian distribution with a nonvanishing population correlation coefficient ρ, the Fisher transformation can be used to estimate the confidence interval for ρ from the measured value r. The distribution of F,

$$F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \tag{3.104}$$

approximately follows a Gaussian distribution with the mean $\mu_F = F(\rho)$ and a standard deviation $\sigma_F = (N-3)^{-1/2}$. For the above sample with $N = 10$ and $r = 0.72$, this approximate approach gives a significance level of 0.8% when $\rho = 0$ (instead of the exact value of 1%).

Pearson's correlation coefficient has two main deficiencies. First, the measurement errors for $\{x_i\}$ and $\{y_i\}$ are not used. As they become very large, the significance of any given value of r will decrease (if errors are much larger than the actual ranges of data values, the evidence for correlation vanishes). The case of nonnegligible errors is discussed in chapter 8. Second, Pearson's correlation coefficient is sensitive to Gaussian outliers (recall the discussion in the previous section) and, more generally, the distribution of r does not follow Student's t distribution if $\{x_i\}$ and $\{y_i\}$ are not drawn from a bivariate Gaussian distribution. In such cases, nonparametric correlation tests, discussed next, are a better option.

3.6.1. Nonparametric Correlation Tests

The two best known nonparametric (distribution-free) correlation tests are based on Spearman's correlation coefficient and Kendall's correlation coefficient. These two correlation coefficients are themselves strongly correlated ($r = 0.98$ in the case of Gaussian distributions), though their magnitude is not equal (see below). Traditionally, Spearman's correlation coefficient is used more often because it is easier to compute, but Kendall's correlation coefficient is gaining popularity because it approaches normality faster than Spearman's correlation coefficient.

Both Spearman's and Kendall's correlation coefficients are based on the concept of ranks. To get ranks, sort a data set $\{x_i\}$ in ascending order. The index i of a value x_i in the sorted data set is its rank, R_i^x. Since most data sets of interest

here involve continuous variables, we will ignore the possibility that some values could be equal, but there are also methods to handle ties in rank (see Lup93). The main advantage of ranks is that their distribution is known: each value $(1, \ldots, N)$ occurs exactly once and this fact can be used to derive useful results such as

$$\sum_{i=1}^{N} R_i = \frac{N\,(N+1)}{2} \tag{3.105}$$

and

$$\sum_{i=1}^{N} (R_i)^2 = \frac{N\,(N+1)\,(2N+1)}{6}. \tag{3.106}$$

Spearman's correlation coefficient is defined analogously to Pearson's coefficient, with ranks used instead of the actual data values,

$$r_S = \frac{\sum_{i=1}^{N} (R_i^x - \overline{R^x})(R_i^y - \overline{R^y})}{\sqrt{\sum_{i=1}^{N} (R_i^x - \overline{R^x})^2}\sqrt{\sum_{i=1}^{N} (R_i^y - \overline{R^y})^2}}, \tag{3.107}$$

or alternatively (see Lup93 for a brief derivation),

$$r_S = 1 - \frac{6}{N\,(N^2-1)} \sum_{i=1}^{N} \left(R_i^x - R_i^y\right)^2. \tag{3.108}$$

The distribution of r_S for uncorrelated variables is the same as given by eq. 3.103, with Pearson's r replaced by r_S.

Kendall's correlation coefficient is based on a comparison of two ranks, and does not use their actual difference (i.e., the difference $R_i^x - R_i^y$ in expression above). If $\{x_i\}$ and $\{y_i\}$ are not correlated, a comparison of two pairs of values j and k, defined by $R_j^x = R_j^y$ and $R_k^x = R_k^y$ will produce similar numbers of *concordant pairs*, defined by $(x_j - x_k)(y_j - y_k) > 0$, and *discordant pairs*, defined by $(x_j - x_k)(y_j - y_k) < 0$. The condition for concordant pairs corresponds to requiring that *both* differences have the same sign, and opposite signs for discordant pairs. For a perfect correlation (anticorrelation), all $N(N-1)/2$ possible pairs will be concordant (discordant).

Hence, to get Kendall's correlation coefficient, τ, count the number of concordant pairs, N_c, and the number of discordant pairs, N_d,

$$\tau = 2\,\frac{N_c - N_d}{N(N-1)} \tag{3.109}$$

(note that $-1 \leq \tau \leq 1$). Kendall's τ can be interpreted as the probability that the two data sets are in the same order minus the probability that they are not in the same order.

For small N, the significance of τ can be found in tabulated form. When $N > 10$, the distribution of Kendall's τ, for the case of *no correlation*, can be approximated as

a Gaussian with $\mu = 0$ and width

$$\sigma_\tau = \left[\frac{2(2N+5)}{9N(N-1)}\right]^{1/2}. \tag{3.110}$$

This expression can be used to find a significance level corresponding to a given τ, that is, the probability that such a large value would arise by chance in the case of no correlation. Note, however, that Kendall's τ is not an estimator of ρ in the general case.

When $\{x_i\}$ and $\{y_i\}$ are correlated with a true correlation coefficient ρ, then the distributions of measured Spearman's and Kendall's correlation coefficients become harder to describe. It can be shown that for a bivariate Gaussian distribution of x and y with a correlation coefficient ρ, the expectation value for Kendall's τ is

$$\overline{\tau} = \frac{2}{\pi}\sin^{-1}(\rho) \tag{3.111}$$

(see [7] for a derivation, and for a more general expression for τ in the presence of noise). Note that τ offers an unbiased estimator of the population value, while r_S does not (see Lup93). In practice, a good method for placing a confidence estimate on the measured correlation coefficient is the bootstrap method (see §4.5). An example, shown in figure 3.24 compares the distribution of Pearson's, Spearman's, and Kendall's correlation coefficients for the sample shown in figure 3.23. As is evident, Pearson's correlation coefficient is *very sensitive* to outliers!

The efficiency of Kendall's τ relative to Pearson's correlation coefficient for a bivariate Gaussian distribution is greater than 90%, and can exceed it by large factors for non-Gaussian distributions (the method of so-called normal scores can be used to raise the efficiency to 100% in the case of a Gaussian distribution). Therefore, Kendall's τ is a good general choice for measuring the correlation of any two data sets.

The computation of N_c and N_d needed for Kendall's τ by direct evaluation of $(x_j - x_k)(y_j - y_k)$ is an $\mathcal{O}(N^2)$ algorithm. In the case of large samples, more sophisticated $\mathcal{O}(N \log N)$ algorithms are available in the literature (e.g., [1]).

3.7. Random Number Generation for Arbitrary Distributions

The distributions in `scipy.stats.distributions` each have a method called `rvs`, which implements a pseudorandom sample from the distribution (see examples in the above sections). In addition, the module `numpy.random` implements samplers for a number of distributions. For example, to select five random integers between 0 and 10:

```
>>> import numpy as np
>>> np.random.random_integers(0, 10, 5)
array([7, 5, 1, 1, 6])
```

For a full list of available distributions, see the documentation of `numpy.random` and of `scipy.stats`.

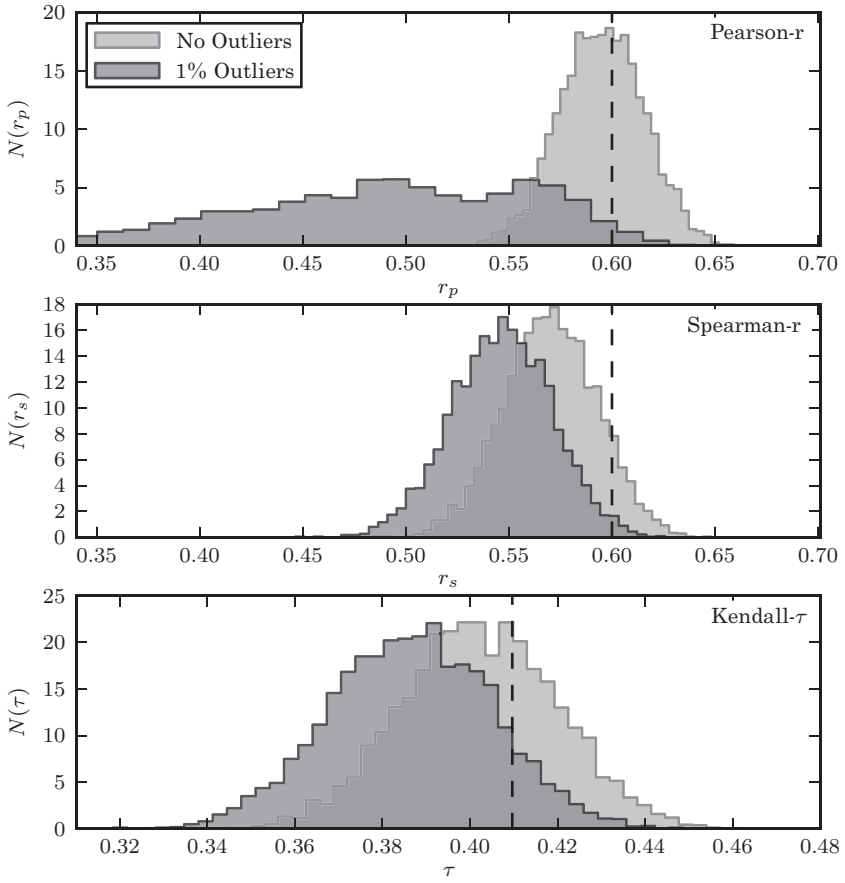

Figure 3.24. Bootstrap estimates of the distribution of Pearson's, Spearman's, and Kendall's correlation coefficients based on 2000 resamplings of the 1000 points shown in figure 3.23. The true values are shown by the dashed lines. It is clear that Pearson's correlation coefficient is not robust to contamination.

Numerical simulations of the measurement process are often the only way to understand complicated selection effects and resulting biases. These approaches are often called Monte Carlo simulations (or modeling) and the resulting artificial (as opposed to real measurements) samples are called Monte Carlo or mock samples. Monte Carlo simulations require a sample drawn from a specified distribution function, such as the analytic examples introduced earlier in this chapter, or given as a lookup table. The simplest case is the uniform distribution function (see eq. 3.39), and it is implemented in practically all programming languages. For example, module `random` in Python returns a random (really pseudorandom since computers are deterministic creatures) floating-point number greater than or equal to 0 and less than 1, called a uniform deviate. The `random` submodule of NumPy provides some more sophisticated random number generation, and can be much faster than the random number generation built into Python, especially when generating large random arrays.

When "random" is used without a qualification, it usually means a uniform deviate. The mathematical background of such random number generators (and

pitfalls associated with specific implementation schemes, including strong correlation between successive values) is concisely discussed in NumRec. Both the Python and NumPy random number generators are based on the Mersenne twister algorithm [4], which is one of the most extensively tested random number generators available. Although many distribution functions are already implemented in Python (in the `random` module) and in NumPy and SciPy (in the `numpy.random` and `scipy.stats` modules), it is often useful to know how to use a uniform deviate generator to generate a simulated (mock) sample drawn from an arbitrary distribution.

In the one-dimensional case, the solution is exceedingly simple and is called the transformation method. Given a differential distribution function $f(x)$, its cumulative distribution function $F(x)$ given by eq. 1.1 can be used to choose a specific value of x as follows. First use a uniform deviate generator to choose a value $0 \leq y \leq 1$, and then choose x such that $F(x) = y$. If $f(x)$ is hard to integrate, or given in a tabular form, or $F(x)$ is hard to invert, an appropriate numerical integration scheme can be used to produce a lookup table for $F(x)$. An example of "cloning" 100,000 data values following the same distribution as 10,000 "measured" values using table interpolation is given in figure 3.25. This particular implementation uses a cubic spline interpolation to approximate the inverse of the observed cumulative distribution $F(x)$. Though slightly more involved, this approach is much faster than the simple selection/rejection method (see NumRec for details). Unfortunately, this rank-based approach cannot be extended to higher dimensions. We will return to the subject of cloning a general multidimensional distribution in §6.3.2.

In multidimensional cases, and when the distribution is separable (i.e., it is equal to the product of independent one-dimensional distributions, e.g., as given for the two-dimensional case by eq. 3.6), one can generate the distribution of each random deviate using a one-dimensional prescription. When the multidimensional distribution is not separable, one needs to consider marginal distributions. For example, in a two-dimensional case $h(x, y)$, one would first draw the value of x using the marginal distribution given by eq. 3.77. Given this x, say x_o, the value of y, say y_o, would be generated using the properly normalized one-dimensional cumulative conditional probability distribution in the y direction,

$$H(y|x_o) = \frac{\int_{-\infty}^{y} h(x_o, y')\, dy'}{\int_{-\infty}^{\infty} h(x_o, y')\, dy'}. \tag{3.112}$$

In higher dimensions, x_o and y_o would be kept fixed, and the properly normalized cumulative distributions of other variables would be used to generate their values.

In the special case of multivariate Gaussian distributions (see §3.5), mock samples can be simply generated in the space of principal axes, and then the values can be "rotated" to the appropriate coordinate system (recall the discussion in §3.5.2). For example, two independent sets of values η_1 and η_2 can be drawn from an $\mathcal{N}(0, 1)$ distribution, and then x and y coordinates can be obtained using the transformations (cf. eq. 3.88)

$$x = \mu_x + \eta_1 \sigma_1 \cos\alpha - \eta_2 \sigma_2 \sin\alpha \tag{3.113}$$

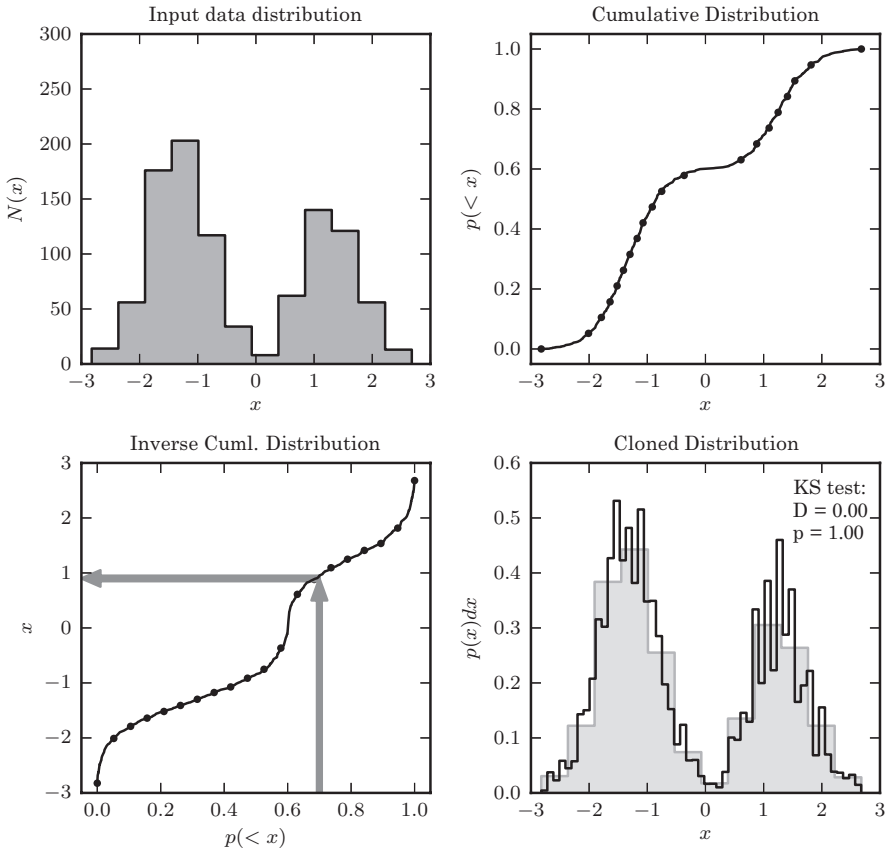

Figure 3.25. A demonstration of how to empirically clone a distribution, using a spline interpolation to approximate the inverse of the observed cumulative distribution. This allows us to nonparametrically select new random samples approximating an observed distribution. First the list of points is sorted, and the rank of each point is used to approximate the cumulative distribution (upper right). Flipping the axes gives the inverse cumulative distribution on a regular grid (lower left). After performing a cubic spline fit to the inverse distribution, a uniformly sampled x value maps to a y value which approximates the observed pdf. The lower-right panel shows the result. The K-S test (see §4.7.2) indicates that the samples are consistent with being drawn from the same distribution. This method, while fast and effective, cannot be easily extended to multiple dimensions.

and

$$y = \mu_y + \eta_1 \sigma_1 \sin\alpha + \eta_2 \sigma_2 \cos\alpha. \tag{3.114}$$

The generalization to higher dimensions is discussed in §3.5.4.

The cloning of an arbitrary high-dimensional distribution is possible if one can sufficiently model the density of the generating distribution. We will return to this problem within the context of density estimation routines: see §6.3.2.

References

[1] Christensen, D. (2005). Fast algorithms for the calculation of Kendall's τ. *Computational Statistics 20*, 51–62.

[2] Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *Am. Jour. Phys. 14*, 1–13.

[3] Gardner, M. (1959). Mathematical games. *Scientific American 201*, 180–182.

[4] Matsumoto, M. and T. Nishimura (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul. 8*(1), 3–30.

[5] Shevlyakov, G. and P. Smirnov (2011). Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics 40*, 147–156.

[6] Vos Savant, M. (1990). Game show problem. *Parade Magazine*, 16.

[7] Xu, W., Y. Hou, Y. S. Hung, and Y. Zou (2010). Comparison of Spearman's rho and Kendall's tau in normal and contaminated normal models. *ArXiv:cs/1011.2009*.